Université des Sciences et de la
Technologie Houari Boumediene

Data Mining

Rapport du mini projet
Partie 1 : Exploration des données

Rédaction:
MOULAI Hassina Safaa
Matricule : 201400007564
HOUACINE Naila Aziza
Matricule : 201400007594
M2 SII Groupe:3

Professeur
Mme. BABA ALI

November 5, 2018

Contents

List of Figures

Chapter 1

Exploration des données

1.1 La bibliothèque WEKA [1]

Nous avons eu recours à la bibliothèque **Weka**, en incluant *weka.jar* et *weka-src.jar*. Nous avons utilisé uniquement trois (3) classes qui sont :

- **Attribute:** Elle modélise un attribut qui peut être de trois types : [numérique, nominal, String],parmi les méthodes utilisées: enumerateValues(), indexOfValue(String), isNominal(), isNumeric(), numValues(), value(position).
- **Instance:** Elle modélise une instance d'un Data Set, celle-ci ayant une valeur ou un "?" pour chaque attribut, citons les méthodes utilisées : attribute(position), enumerateAttributes(), isMissing(Attribute), value(Attribute), setValue(position, double).
- **Instances:** Elle représente l'ensemble de toutes les instances du Data Set, l'on à utilisé les méthodes suivante : attributeStats(position), enumerateInstances(), instance(position), numInstances().

1.2 Affichage d'un DataSet

Une liste de Data Set est accessible à partir de notre application, pour cela une liste défilante nommé **"Datasets"** est disponible sur la barre des menus.

Une fois un Data Set sélectionne, l'on doit appuyer sur le bouton **"Afficher"** afin qu'un tableau (instance , attribut) ne s'affiche sur la gauche de notre interface, en plus de quelques informations tel que le nombre d'instance et nombre d'attributs.

Mais aussi dans le cas d'un Data Set sans valeurs manquantes, une description et liste de caractéristique de chaque attribut apparaitra sur la partie droite de l'interface.

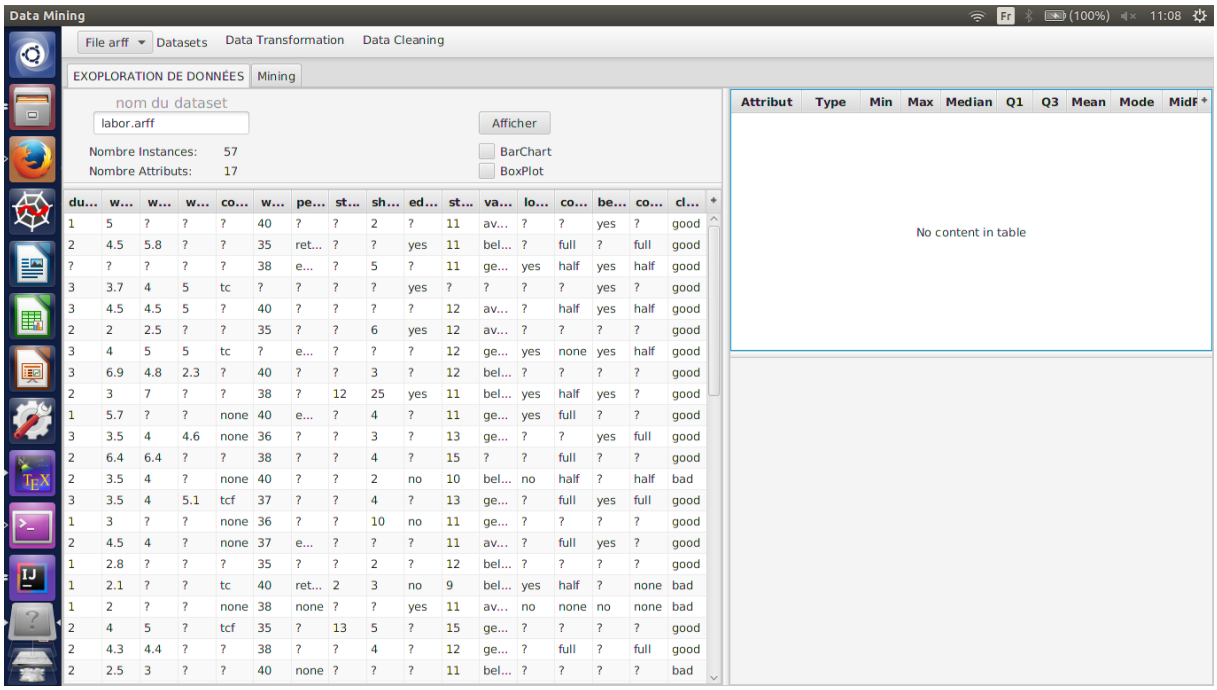

Figure 1.1: Interface de l'application affichant le Data Set [labor.arff] .

1.3 Remplacer les valeurs manquantes

1.3.1 but

Remplacer les valeurs manquantes a pour principale but de redonner une certaine qualité au données, et éviter de tomber dans des cas d'inconsistance et d'incohérence par la suite dans le processus d'extraction de connaissances .

1.3.2 Méthode : entièrement implémenter (Orienté classe)

Valeurs numériques :

Cette première méthode a pour principe de étant donnée une instance I ayant un attribut Ai manquant et appartenant à une classe C , on devra alors remplacer la valeur manquante de Ai par la Moyenne des valeurs des instances appartenant à la Classe C ayant une valeur dans l'attribut Ai (non manquante).

Implémentation:

Pour l'implémentation de cette méthode on a choisit de construire un dictionnaire (index) de façon à avoir: la clé c'est la Classe, les valeurs sont les instances I ayant une valeur (non manquante) dans l'attribut Ai.

exemple: Dictionnaire pour l'attribut DURATION (numeric)

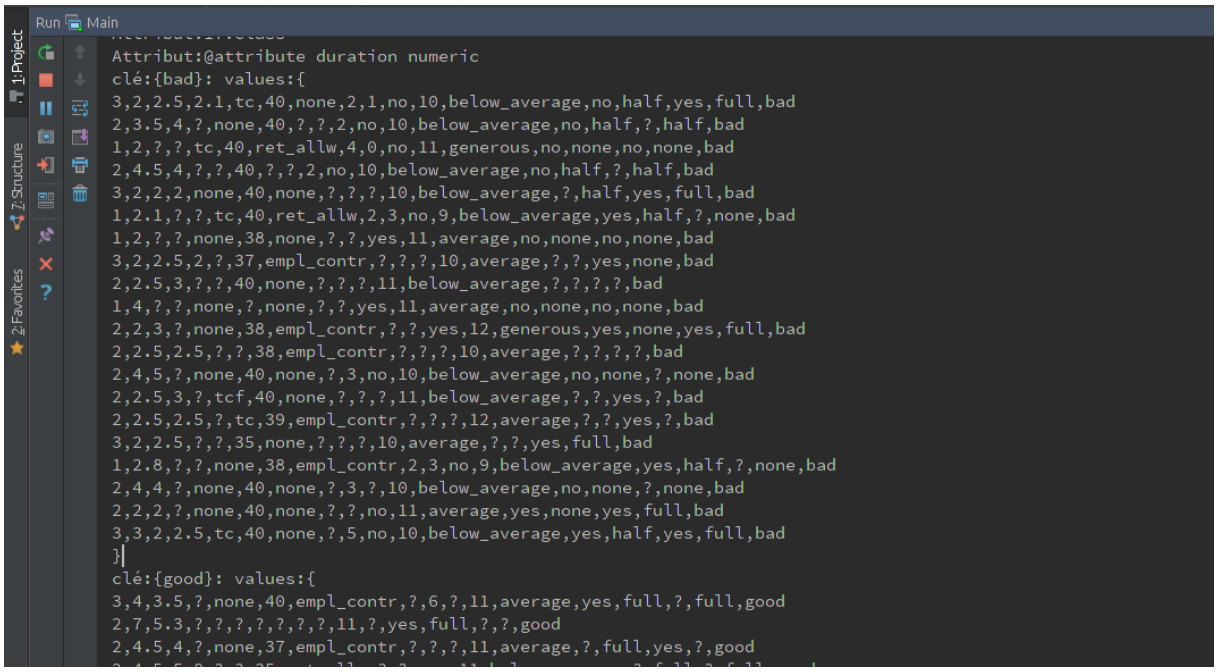

Figure 1.2: Dictionnaire par Classe regroupant les instances dont la valeur de l'attribut DURATION existe, Data Set [labor.arff] .

exemple: Dictionnaire pour l'attribut WAGE INCREASED EACH YEAR (numeric)

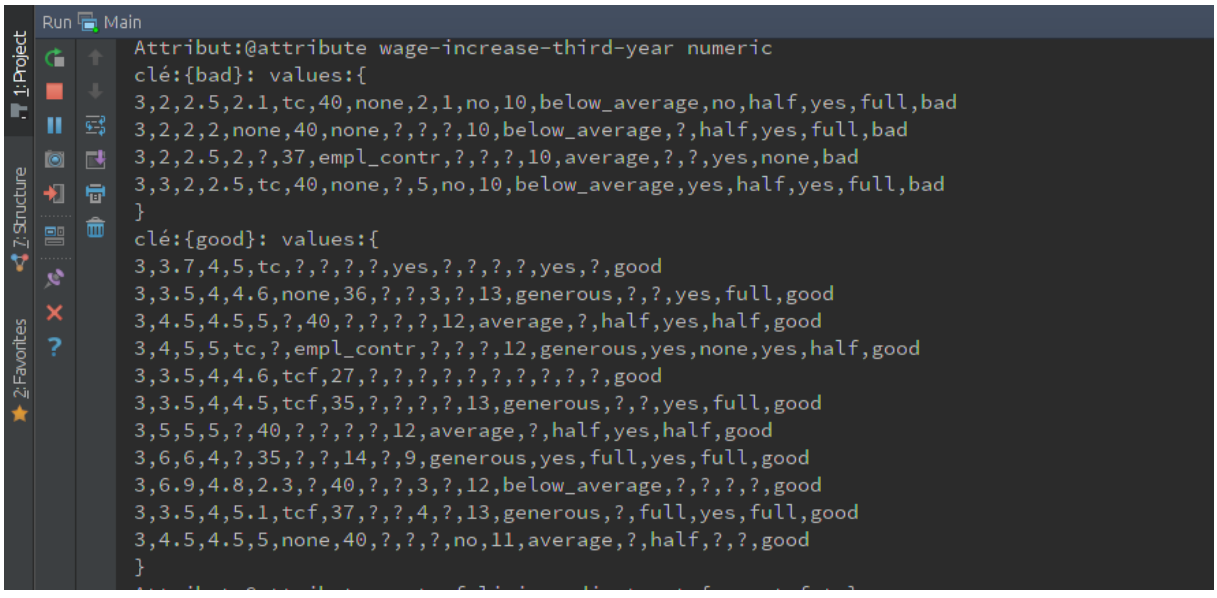

Figure 1.3: Dictionnaire par Classe regroupant les instances dont la valeur de l'attribut WAGE INCREASED EACH YEAR existe, Data Set [labor.arff].

Valeurs nominales: Pour les valeurs nominal on choisit d'exploiter notre structure de dictionnaire et pour une instance I assigner pour chaque attribut manquant Ai appartenant à une classe C, le MODE de la Classe à laquelle l'instance Appartient (Classe C).

exemple: Dictionnaire pour l'attribut contribution-to-dental-plan

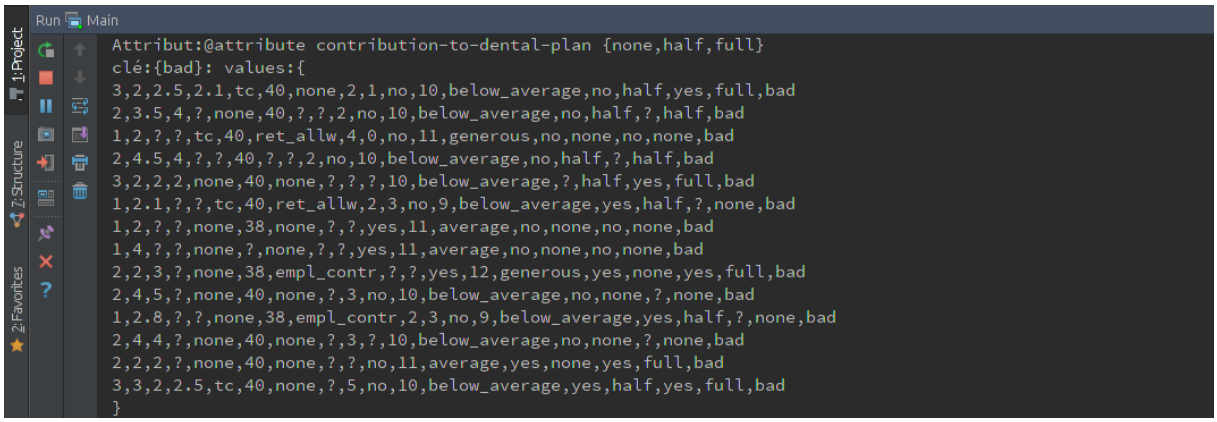

Figure 1.4: Dictionnaire pour la classe bad regroupant les instances dont la valeur de l'attribut contribution-to-dental-plan existe, Data Set [labor.arff].

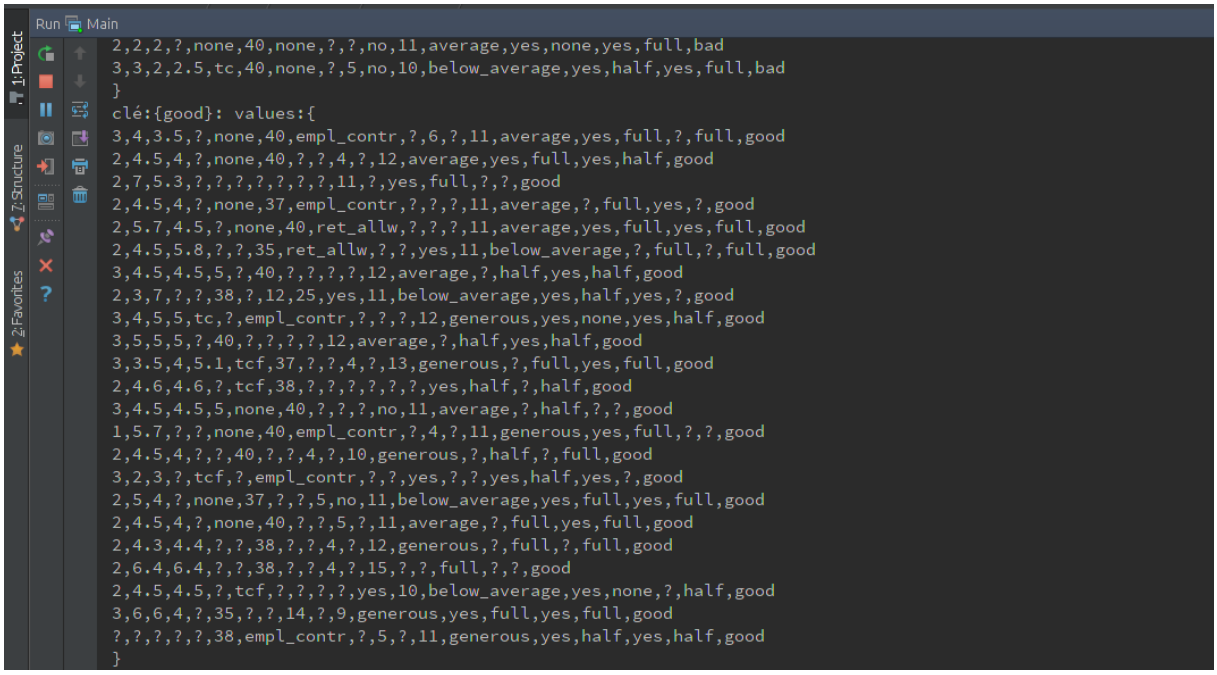

Figure 1.5: Dictionnaire pour la classe good regroupant les instances dont la valeur de l'attribut contribution-to-dental-plan existe, Data Set [labor.arff].

1.3.3 Résultat

Le Data Set avant de remplacer les valeurs manquantes est comme suit:

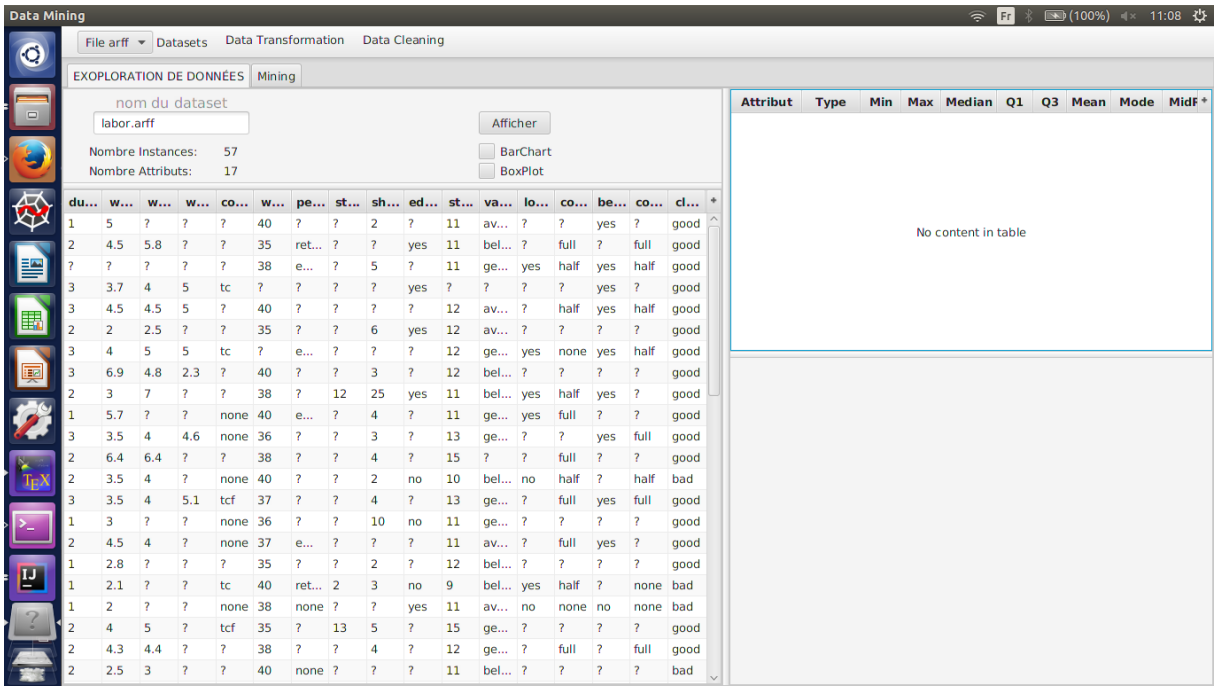

Figure 1.6: Data Set [labor.arff] avant traitement des valeurs manquantes.

L'on remarque les points d'interrogation "?" représentant les valeurs manquantes aussi bien pour les attributs nominaux que numériques.

Après avoir sélectionné dans la barre des menus l'option **Data Cleaning** > **Missing value class** l'on obtient le résultat suivant:

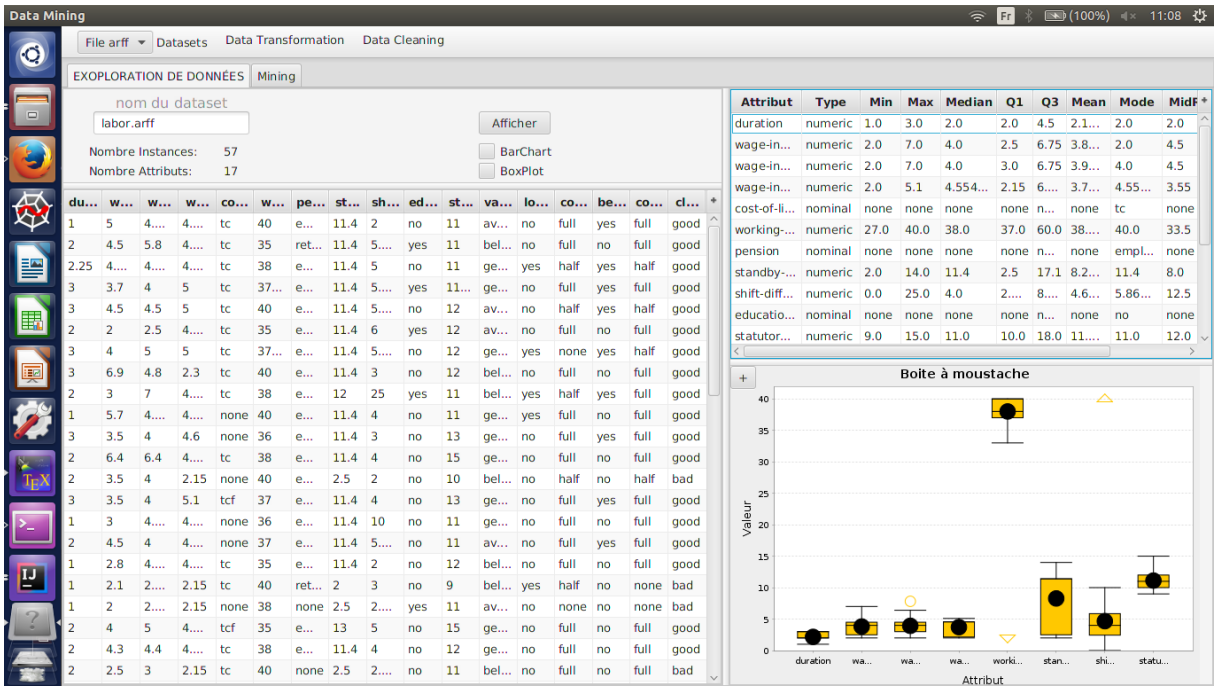

Figure 1.7: Data Set [labor.arff] après traitement des valeurs manquantes.

1.4 Discrétisation

1.4.1 But[2]

Discrétiser un attribut numérique (à valeurs quantitatives) c'est transformer l'ensemble des valeurs réels d'un attribut en un autre ensemble réduit d'intervalles ou de classes représentatifs du premier ensemble. On dis aussi "réaliser un découpage en classes".
Lorsque le nombre de valeur d'un attribut est énorme, l'affichage des histogrammes pour l'étude des caractéristiques de cet attribut est illisible, ou les informations que l'on en tire ne sont pas pertinentes, c'est pour cela que l'on à recours à la discrétisation.

1.4.2 Méthode

Choix du nombre d'intervalles

Il existe quelques formules de moindre complexité pour déterminer à l'aveugle le nombre de classes à partir de:
N : le nombre total de données.
Min : le minimum des données.
Max : le maximum des données.
IQR : l'écart inter-quartiles (paramètres de la dispersion).

Nous avons choisi la fonction de **Freedman-Diaconis** qui est la suivante:

$$NbrIntervalle = \frac{Max - Min}{2 * IQR * N^{-\frac{1}{3}}}$$

Création des intervalles/classes

Parmi plusieurs méthodes existantes nous avons opté pour **La méthode des amplitudes** c'est à dire à tailles (amplitudes) égales.
Tel que la taille de chaque intervalle est égale à:

$$Amplitude = \frac{(Max - Min)}{NbrIntervalle}$$

Ainsi en le 1^{er} intervalle $= [Min : Min + Amplitude]$,
le suivant $= [Min + Amplitude : Min + 2 * Amplitude]$, ...

Remplacement de chaque valeur par la borne supérieur de l'intervalle au quel elle appartient

Ainsi pour chaque attribut , nous parcourons toutes les valeurs de celui-ci et pour chacune de ses valeur nous vérifions à quel intervalle "I" elle appartient, puis nous remplaçons cette valeur par la borne supérieure de I.

PS: L'attribut étant de type numérique et afin de ne pas dénaturer la dataset, nous avons retenu uniquement la borne supérieur des intervalles pour remplacer les valeurs lors de la discrétisation.

1.4.3 Résultat

Le Data Set avant discrétisation est comme suit:

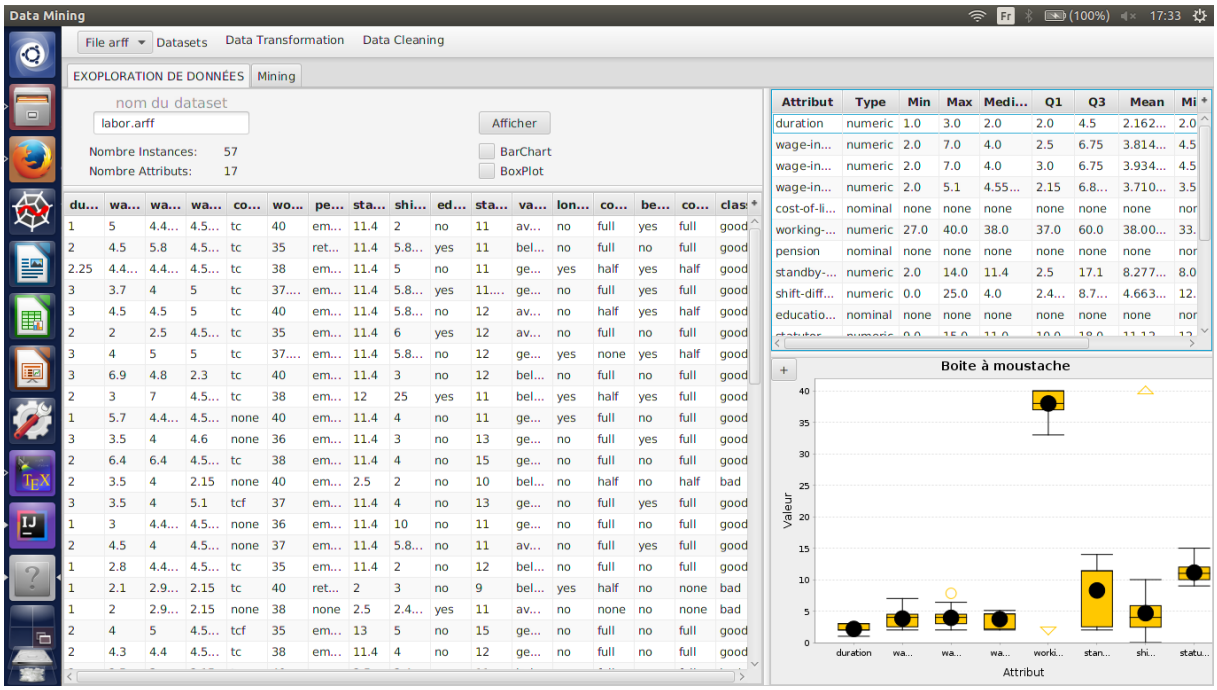

Figure 1.8: Data Set [labor.arff] avant discrétisation.

Après avoir sélectionné dans la barre des menus l'option **Data Transformation** > **Discrétisation** l'on obtient le résultat suivant:

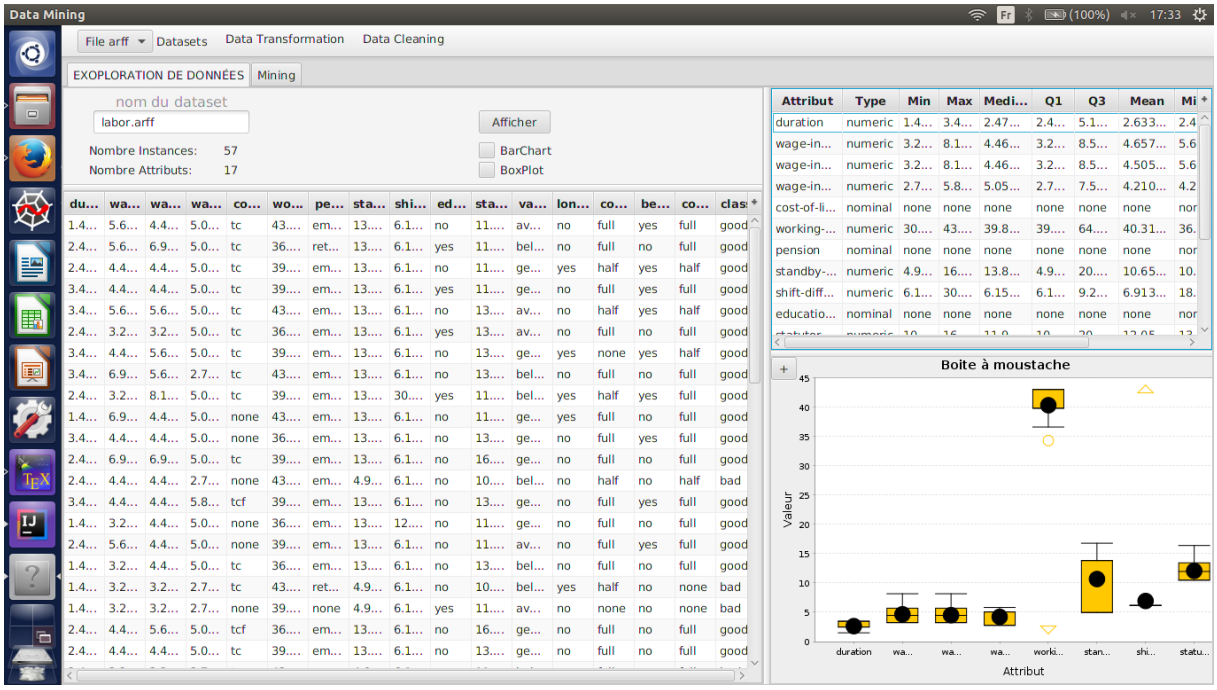

Figure 1.9: Data Set [labor.arff] après discrétisation.

Remarque Nous remarquons ainsi que le degré de similitude entre les instances par rapport aux attributs est plus visible, ainsi nous espérons avoir de meilleurs résultats pour les prochaines étapes.

1.5 Normalisation

1.5.1 But

Pour la visualisation des données, et surtout la comparaison des distributions et caractéristiques des différents attributs d'un Data Set, nous avons recours à la normalisation des données de ce dernier, car les attributs des Data Set proviennent avec des ordres de grandeurs différents.

en plus que les attributs à grande valeur numérique influeront plus les résultats lors des traitements, apprentissage, ...

1.5.2 Méthode

Cette étape est réalisable en alignant tous les attributs sur une même plage de valeur, généralement [0:1], la formule est la suivante:

$$normalisee = \frac{originale - MIN}{MAX - MIN}$$

[MIN,MAX] : intervalle d'origine.
originale : valeur dans l'intervalle d'origine.
normalisee : valeur normalisée dans l'intervalle cible [0,1].

1.5.3 résultat

Le Data Set avant la normalisation est comme suit:

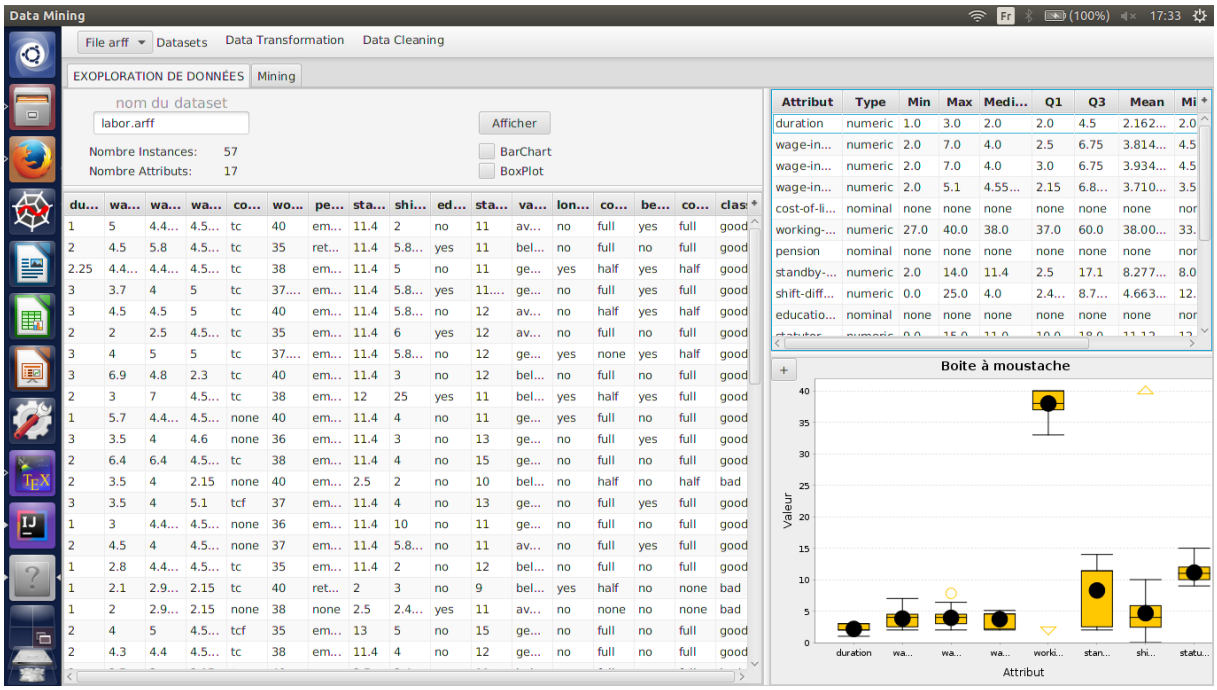

Figure 1.10: Data Set [labor.arff] avant normalisation.

Après avoir sélectionné dans la barre des menus l'option **Data Transformation** > **Normalisation** l'on obtient le résultat suivant:

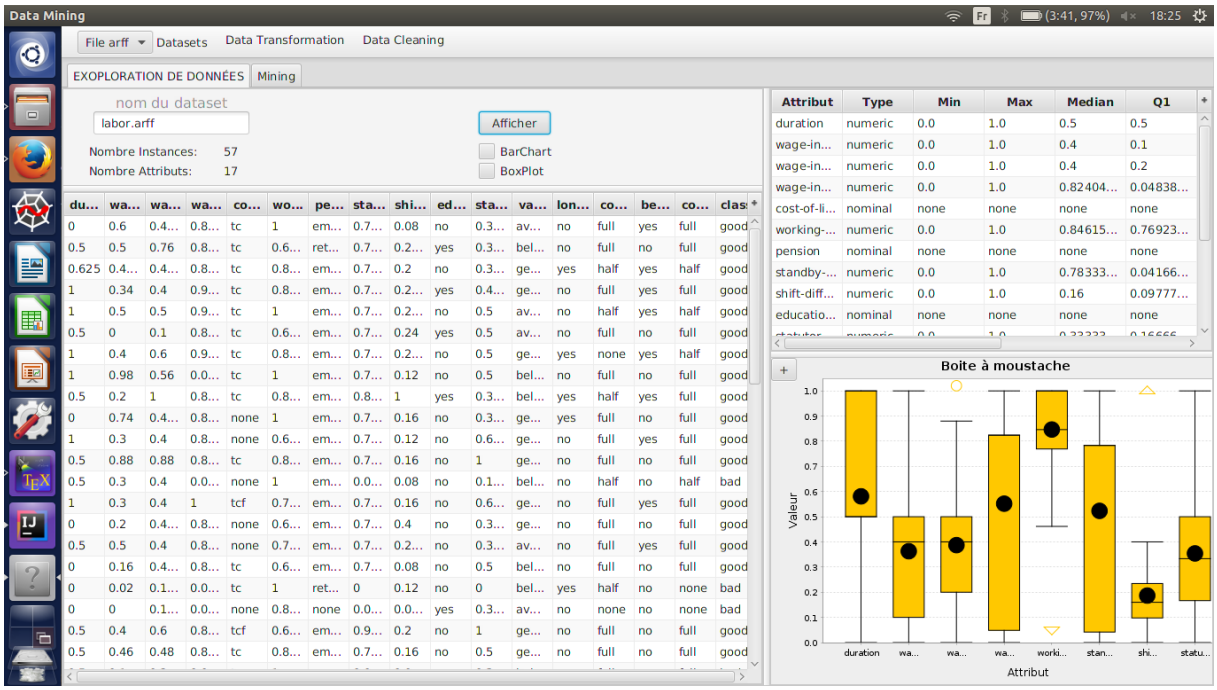

Figure 1.11: Data Set [labor.arff] après normalisation.

Remarque: Il est claire qu'il est plus simple de comparer les attributs une fois uniformisé, comme cela apparait dans l'affichage des boites à moustaches.

1.6 Description du Data Set et de ses attributs

Le Data Set "Labor.arff" choisi pour cette analyse comporte 16 attributs ainsi que la classe. Afin de mener à bien notre étude / analyse de ce Data Set nous aurons recours à deux moyen de visualisation qui sont les Histogramme (BarChart) ainsi que les Boites à moustache (BoxPlot), ce dernier est utilisé uniquement sur les attributs numériques, et donnée par les figures ci-dessous:

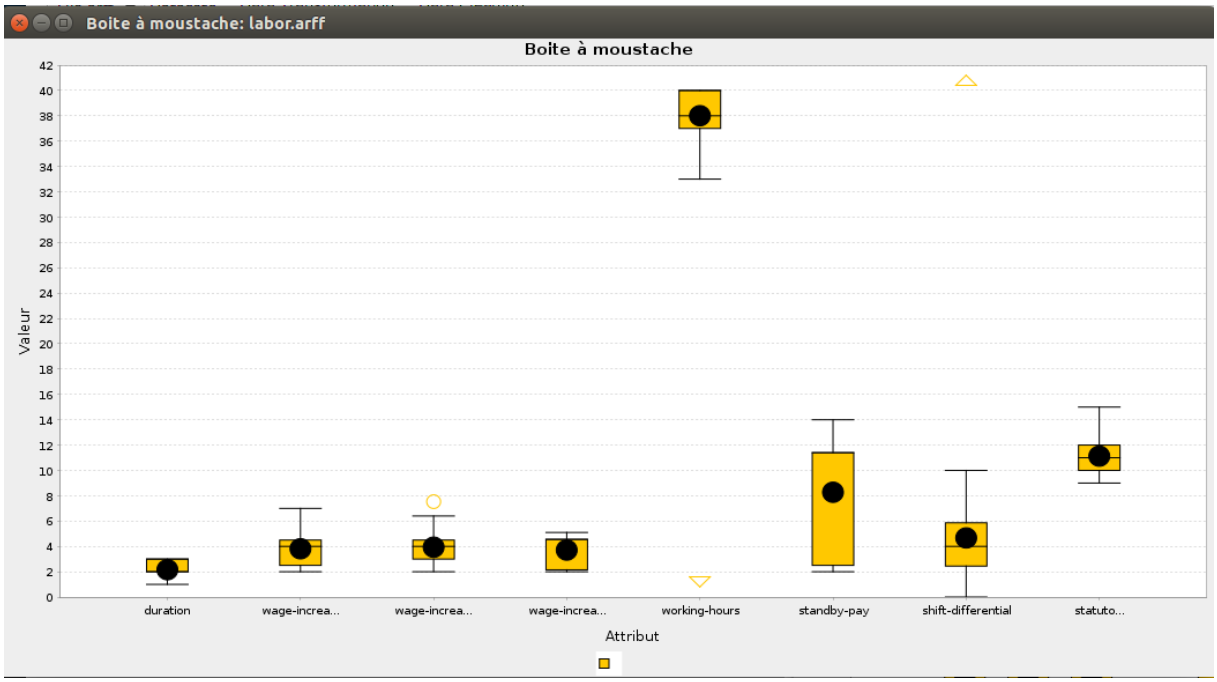

Figure 1.12: Boites à moustaches du Data Set [labor.arff] brute.

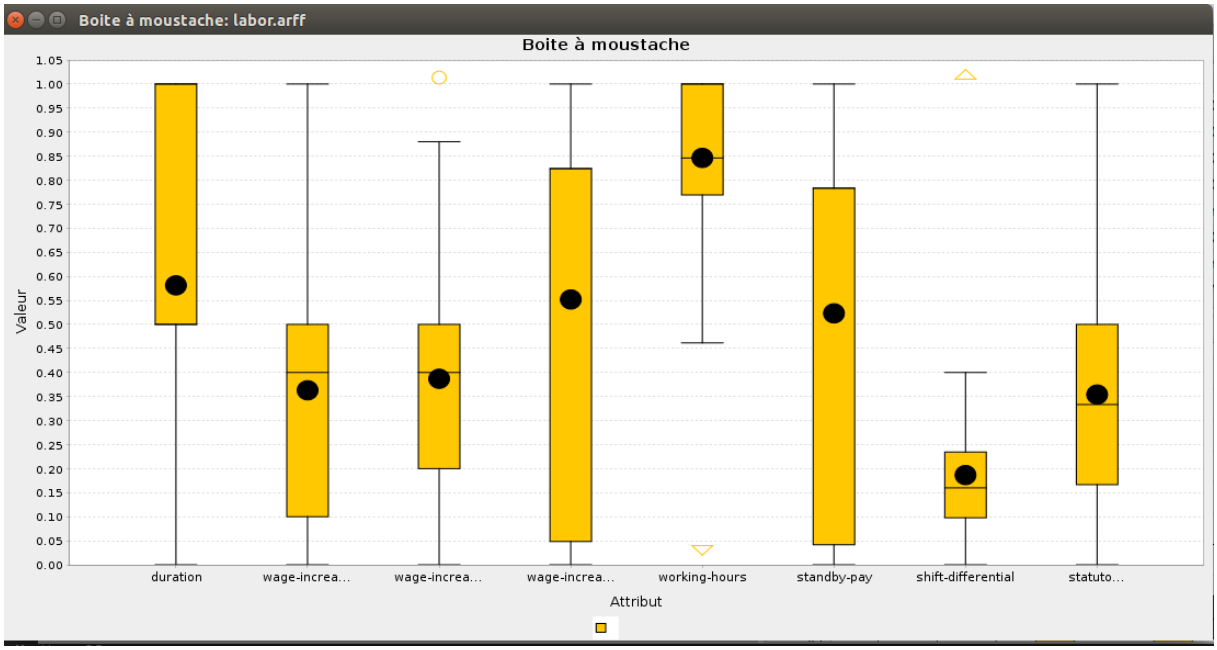

Figure 1.13: Boites à moustaches du Data Set [labor.arff] après normalisation.

Remarque Ces figures seront détaillées lors de la description des attributs.

1.6.1 Attribut 1 : duration

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "durée de l'accord" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
duration	numeric	1.0	3.0	2.0	2.0	4.5	2.16228...	2.0	2.0	Symetric

Figure 1.14: Attribut 1 : duration du Data Set [labor.arff] .

On se basant sur la comparaison les valeurs de *[Mode , Médiane , Moyenne] = (2 , 2 , 2.1)*, nous déduisons la dispersion des données selon cet attribut, tel qu'il est **Symétrique**.

BarChart

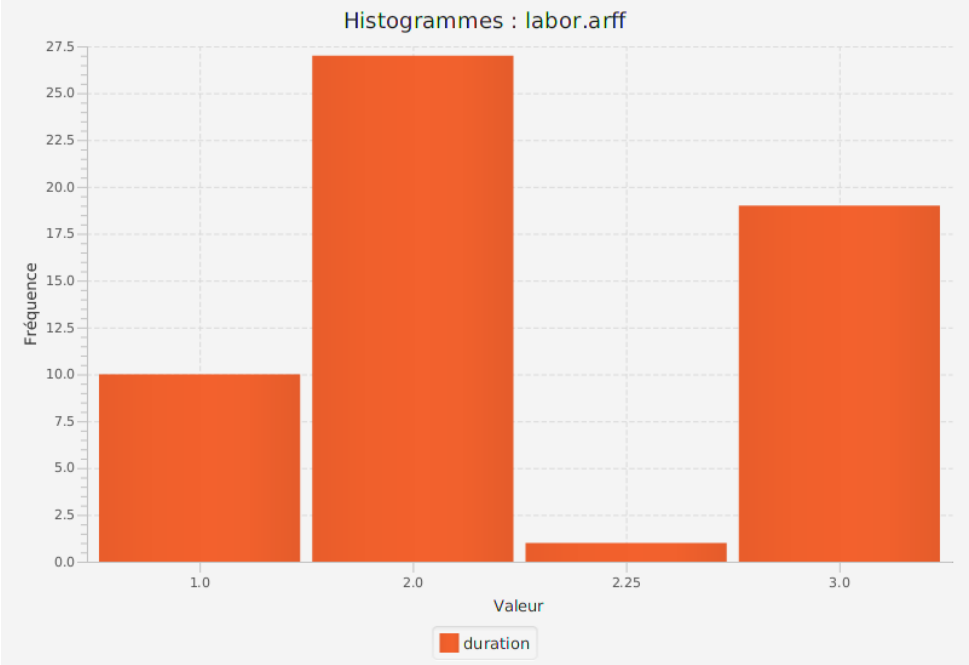

Figure 1.15: Histogramme de duration du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 4 valeurs possibles *(1.0 , 2.0 , 2.25 , 3.0)*. la répartition n'est pas uniforme vu la fréquence d'apparition de la valeur *(2.25)* comparé à celle de la valeur *(2.0)*.

• Il est aussi à noter que cet attribut est **Uni-modal**, tel que mode = *(2.0)*.

• A travers sa boite à moustache de cet attribut nous constatons une répartition pas très uniforme des données, dans le sens ou il y a une concentration des données pour les valeurs supérieurs entre (2.0 , 3.0).

1.6.2 Attribut 2 : wage-increase-first-year

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "augmentation en première année de contrat" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
wage-increase-first-year	numeric	2.0	7.0	4.0	2.5	6.75	3.81437...	4.5	2.0	Positively skewed

Figure 1.16: Attribut 2 : wage-increase-first-year du Data Set [labor.arff] .

On se basant sur la comparaison les valeurs de *[Mode , Médiane , Moyenne] = (2.0 , 4.0 , 3.81)*, nous déduisons la dispersion des données selon cet attribut, tel qu'il est **asymétrique positivement**.

BarChart

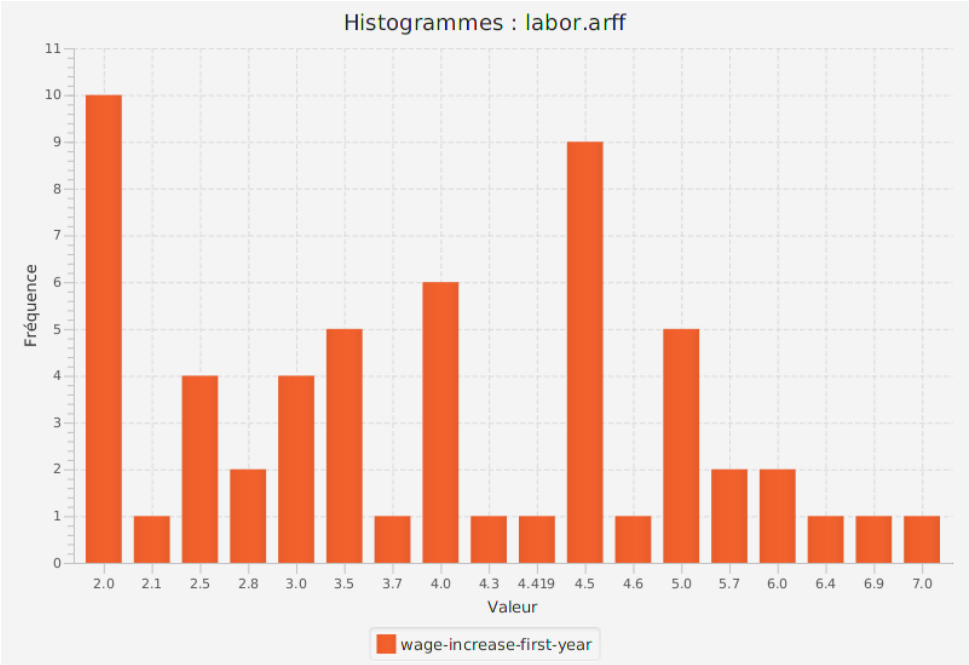

Figure 1.17: Histogramme de wage-increase-first-year du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 18 valeurs possibles *(2.0 , 2.1 , 2.5 , 2.8 , 3.0 , 3.5 , 3.7 , 4.0 , 4.3 , 4.419 , 4.5 , 4.6 , 5.0 , 5.7 , 6.0 , 6.4 , 6.9 , 7.0)*. la répartition n'est pas uniforme vu par exemple : la fréquence d'apparition de la valeur *(7.0)* comparé à celle de la valeur *(2.0)* ou de la valeur *(4.5)*.

• Il est aussi à noter que cet attribut est **Uni-modal**, tel que mode = *(2.0)*.

• A travers sa boite à moustache de cet attribut nous constatons une répartition plus ou moins uniforme des données, mis à part une légère tendance vers les valeurs inférieur.

1.6.3 Attribut 3 : wage-increase-second-year

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "augmentation en deuxième année de contrat" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
wage-increase-second-year	numeric	2.0	7.0	4.0	3.0	6.75	3.93476...	4.5	4.0	Symetric

Figure 1.18: Attribut 3 : wage-increase-second-year du Data Set [labor.arff] .

On se basant sur la comparaison les valeurs de *[Mode , Médiane , Moyenne] = (4.0 , 4.0 , 3.93)*, nous déduisons la dispersion des données selon cet attribut, tel qu'il est **symétrique**.

BarChart

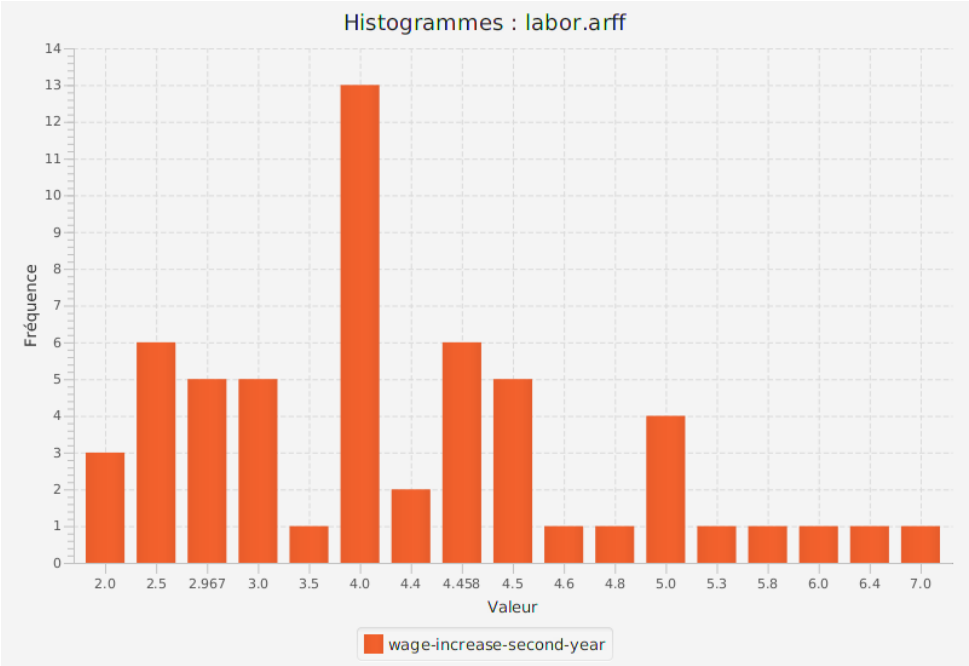

Figure 1.19: Histogramme de wage-increase-second-year du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 16 valeurs possibles *(2.0 , 2.5 , 2.967 , 3.0 , 3.5 , 4.0 , 4.4 , 4.458 , 4.6 , 4.8 , 5.0 , 5.3 , 5.8 , 6.0 , 6.4 , 7.0)*. la répartition n'est pas uniforme vu par exemple : la fréquence d'apparition de la valeur *(7.0)* comparé à celle de la valeur *(4.0)*.

• Il est aussi à noter que cet attribut est **Uni-modal**, tel que mode = *(4.0)*.

• La boite à moustache de cet attribut nous permet de remarquer que la distribution des valeurs n'est pas tout à fait uniforme, tel que plusieurs valeur assez grandes n'apparaissent qu'une fois dans tous le Data Set, ce qui signifie qu'elles peuvent être considérées comme des valeurs aberrantes (out-liers).

1.6.4 Attribut 4 : wage-increase-third-year

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "augmentation en troisième année de contrat" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
wage-increase-third-year	numeric	2.0	5.1	4.55454...	2.15	6.83181...	3.71084...	3.55	4.5545454...	Symetric

Figure 1.20: Attribut 4 : wage-increase-third-year du Data Set [labor.arff] .

On se basant sur la comparaison les valeurs de *[Mode , Médiane , Moyenne] = (4.0 , 4.0 , 3.93)*, nous déduisons la dispersion des données selon cet attribut, tel qu'il est **symétrique**.

BarChart

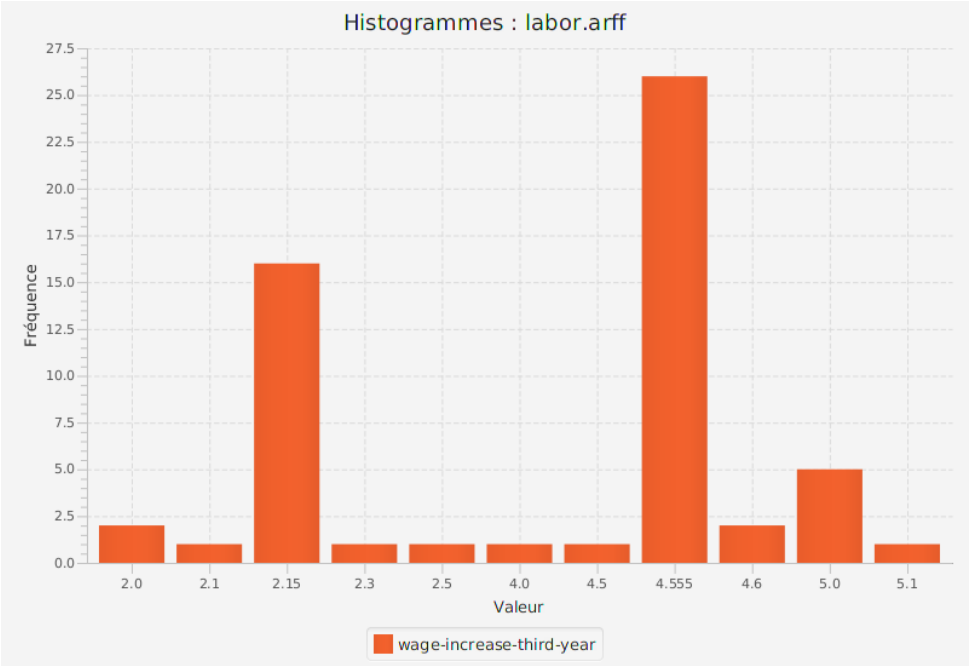

Figure 1.21: Histogramme de wage-increase-third-year du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 11 valeurs possibles *(2.0 , 2.1 , 2.15 , 2.3 , 2.5 , 4 , 4.5 , 4.555 , 4.6 , 5.0 , 5.1)*. la répartition n'est pas uniforme vu par exemple : la fréquence d'apparition de la valeur *(5.1)* comparé à celle de la valeur *(2.15)*.

• Il est aussi à noter que cet attribut est **Uni-modal**, tel que mode = *(4.555)*.

• La boite à moustache de cet attribut nous permet d'apprécier la distribution uniforme de ses valeurs, ainsi aucune valeur aberrante n'y figure.

1.6.5 Attribut 5 : cost-of-living-adjustment

Caractéristiques

Nous résumons les caractéristiques de cet attribut dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
cost-of-living-adjustment	nominal	none	none	none	none	none	none	none	none	none

Figure 1.22: Attribut 5 : cost-of-living-adjustment du Data Set [labor.arff] .

BarChart

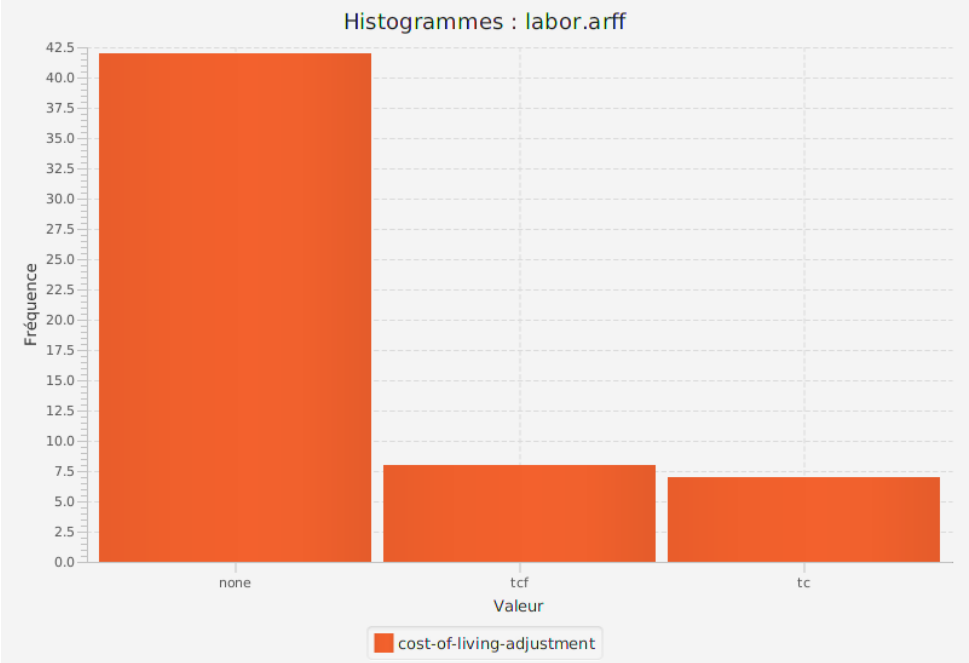

Figure 1.23: Histogramme de cost-of-living-adjustment du Data Set [labor.arff] .

discussion

• On remarque ici facilement que la valeur la plus fréquente est la **"none"** ,avec environ 70% du tout le data set , on retrouve **"tcf","tc"** avec seulement 12% chacun.

• On peut conclure que l'attribut est **Uni-modal** (NOne est le mode et il est unique), de plus la distribution n'est pas uniforme tel que deux valeurs **(none)** prend le dessus sur plus de 70% de tout le Data Set.

1.6.6 Attribut 6 : working-hours

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "nombre d'heure de travail" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
working-hours	numeric	27.0	40.0	38.0	37.0	60.0	38.0023...	33.5	40.0	Negatively skewed

Figure 1.24: Attribut 6 : working-hours du Data Set [labor.arff] .

On se basant sur la comparaison les valeurs de *[Mode , Médiane , Moyenne] = (40.0 , 38.0 , 38.002)*, nous déduisons la dispersion des données selon cet attribut, tel qu'il est légèrement **asymétrique négativement**.

BarChart

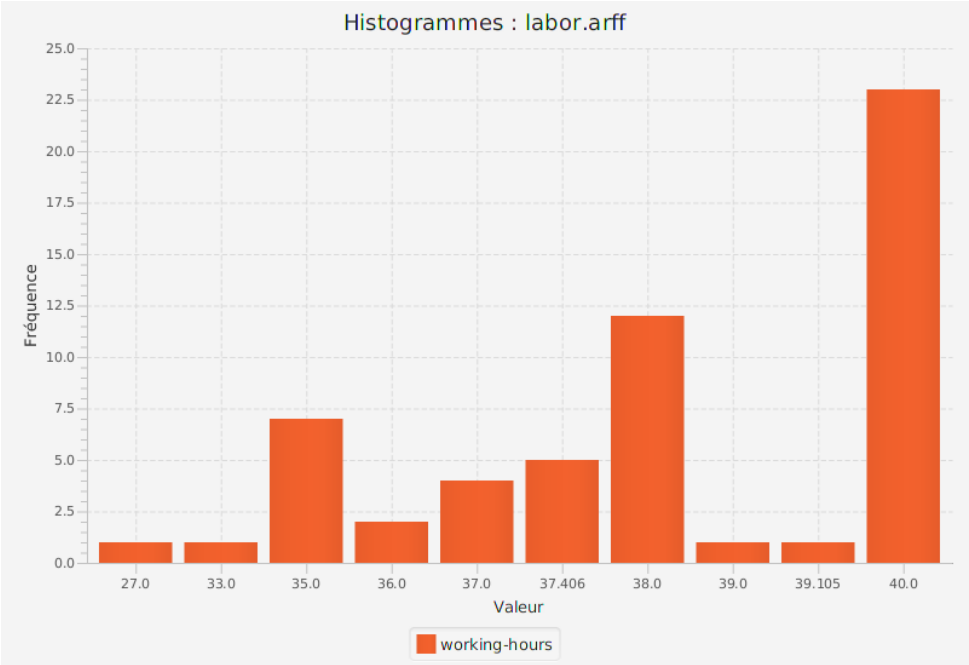

Figure 1.25: Histogramme de working-hours du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 10 valeurs possibles *(27.0 , 33.0 , 35.0 , 36.0 , 37.0 , 37.406 , 38.0 , 39.0 , 39.105 , 40.0)*. la répartition n'est pas uniforme vu par exemple la fréquence d'apparition de la valeur *(40.0)* comparé à celle de la valeur *(27.0)*.

• Il est aussi à noter que cet attribut est **Uni-modal**, tel que l'unique mode = *(40.0)*.

• Aussi à partir de la boite à moustache de cet attribut, nous observons que l'ensemble des données (instances) est concentré entre (37.0 , 40.0) hors que le min est de 27h qui n'apparait qu'une seule fois, d'ailleurs se détail est bien représenté dans la boite à moustache par une flèche orange vers le bas, cela signifie la présence de valeur aberrante (out-lier).

1.6.7 Attribut 7 : pension

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "contribution des employées à la pension" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
pension	nominal	none	none	none	none	none	none	none	empl_contr	none

Figure 1.26: Attribut 7 : pension du Data Set [labor.arff] .

BarChart

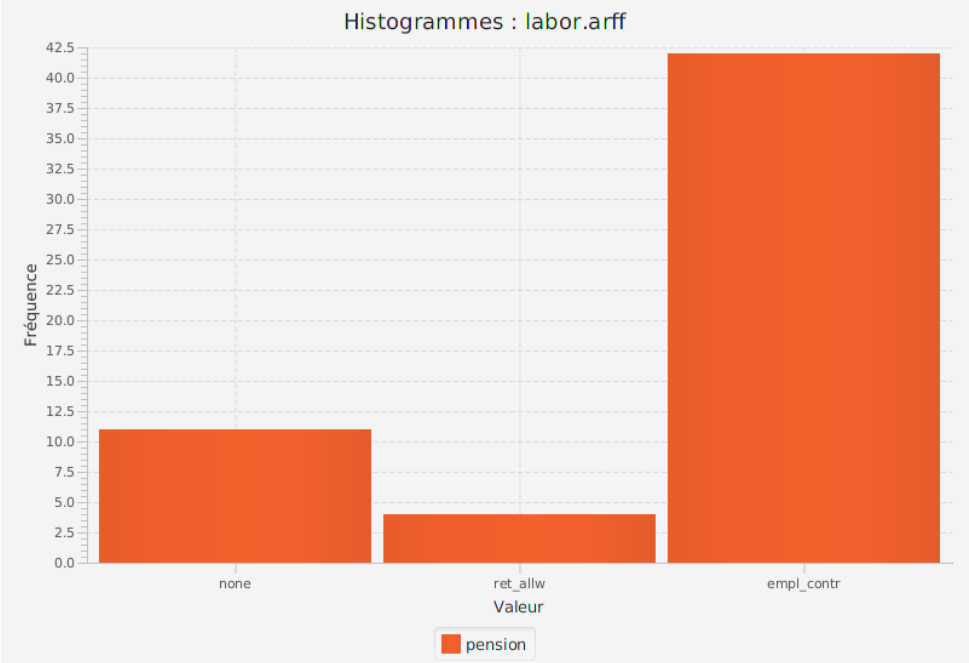

Figure 1.27: Histogramme de pension du Data Set [labor.arff] .

discussion

• Comme vu plus haut l'attribut pension(contribution des employées à la pension) a trois valeurs possible (none,ret_allw(retraité),empl_contr), **empl_contr** étant la plus fréquente avec 75% , suivie de **none** (aucune valeur) à 17% et plus rarement (ret_allw) ayant 4% seulement.

• Pension est alors uni-modal(avec empl_contr) qui contribue le plus au plan pension, et les retraités beaucoup moins avec seulement 4%, ceux qui restent ne contribuent pas du tout au plan pension (none).

1.6.8 Attribut 8 : standby-pay

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "rémunération au repos" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
standby-pay	numeric	2.0	14.0	11.4	2.5	17.1	8.27719...	8.0	11.4	Negatively skewed

Figure 1.28: Attribut 8 : standby-pay du Data Set [labor.arff] .

BarChart

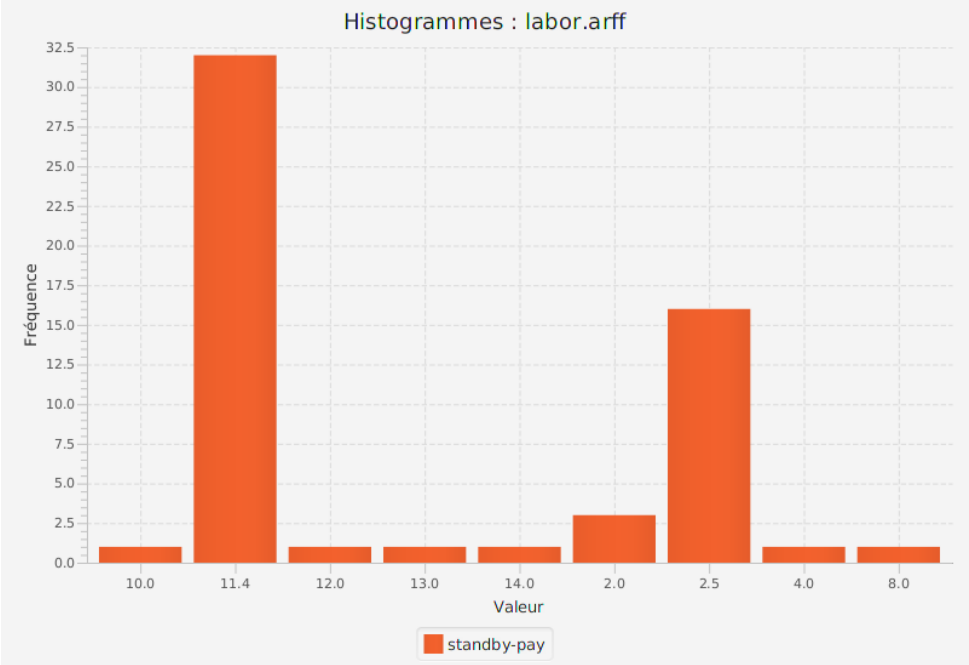

Figure 1.29: Histogramme de standby-pay du Data Set [labor.arff] .

discussion

• L'attribut standby-pay est numérique ,d'aprés l'histogramme de distribution des valeurs on remarque qu'il est uni-modal et la valeur la plus fréquente est de 11.4 suivie de 2.5, à noter que c'est les valeurs calculées auparavant lors des bourrages des valeurs manquantes , on peut déduire alors que avant nettoyage de données l'attribut avait pour mode 2.0 .

• Les valeurs de l'attribut varie entre min 2.0 jusqu'à max 14 ,en observant la boite à moustache on peut voir que la distribution des données est du coté gauche de la médiane ce qui se confirme avec moyenne ¡ à la médiane == mode , le max 14 est ici considéré comme out-lier . ceci nous permet de conclure que Standby-pay est un attribut positivement symétrique .

1.6.9 Attribut 9 : shift-differential

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "supplément pour travail sur les équipes II et III" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
shift-differential	numeric	0.0	25.0	4.0	2.44444...	8.79545...	4.66391...	12.5	5.8636363...	Negatively skewed

Figure 1.30: Attribut 9 : shift-differential du Data Set [labor.arff] .

BarChart

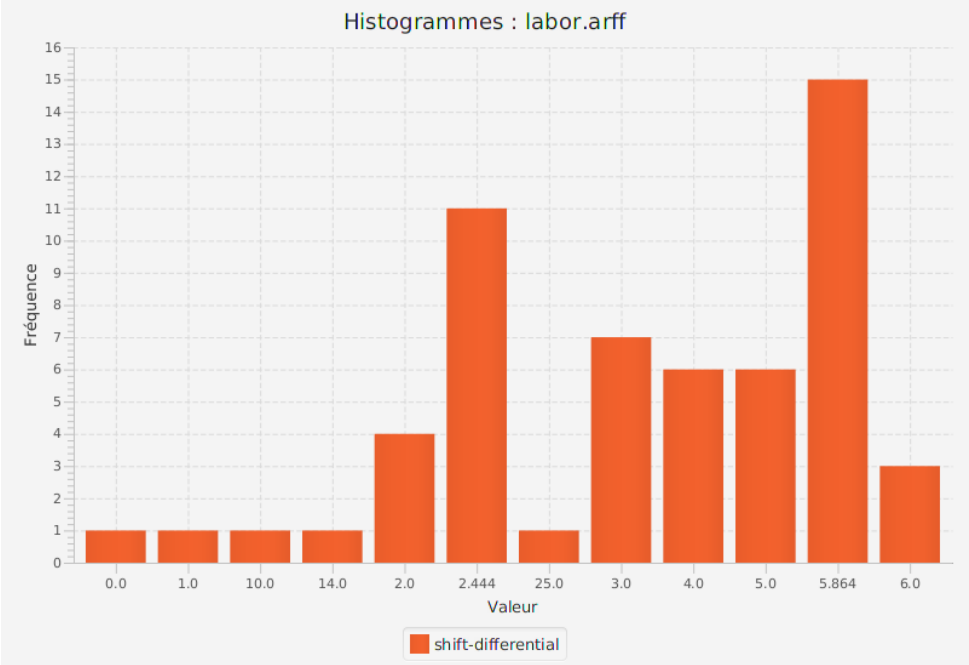

Figure 1.31: Histogramme de shift-differential du Data Set [labor.arff] .

discussion

• L'attribut shift-differential (supplément de travail sur poste 1 , 2 (travail pendant les vacances)...) est numérique ,d'après l'histogramme de distribution des valeurs on remarque facilement qu'il est uni-modal , avec 3 comme unique mode avant nettoyage des données puis 5.8 après bourrages des valeurs manquantes, par contre d'âpres la boite à moustache on observe que la distribution des données (instances) est plutôt loin des min et max qui sont de 0 et 25 respectivement.

25 peut être considéré comme valeur aberrante(out-lier) il n'apparaît qu'une seule fois (une seule instance) et il est très loin de la moyenne et la médiane ainsi que du mode (5.8).

• la médiane étant inférieur à la moyenne et la moyenne au mode donc l'attribut a une distribution qui est légèrement symétrique négativement tel que la plus part des données se concentre après la médiane coté gauche.

1.6.10 Attribut 10 : education-allowance

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "allocation d'éducation" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
education-allowance	nominal	none	none	none	none	none	none	none	no	none

Figure 1.32: Attribut 10 : education-allowance du Data Set [labor.arff] .

BarChart

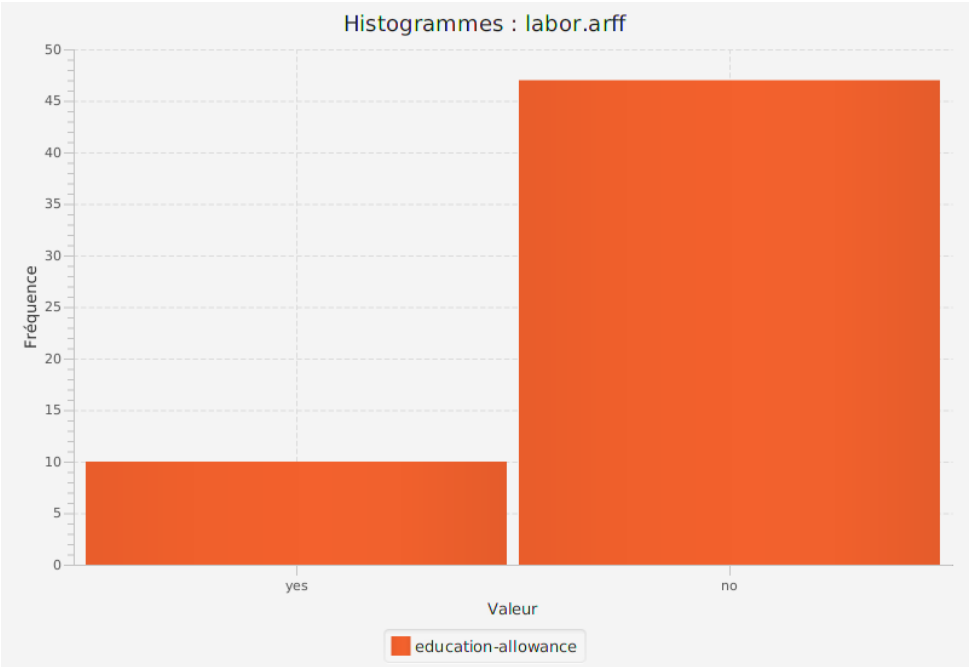

Figure 1.33: Histogramme de education-allowance du Data Set [labor.arff] .

discussion

• L'attribut education-allowance est binaire soit oui, ou non (bourses données par l'employeur à l'employé) .On remarque que dans 84% des on ne donne pas d'education-allowance par contre seulement %16 ont droit à une education-allowance.

• On conclue que "education-allowance" est uni-modal avec la valeur ("NO") la plus fréquente.

1.6.11 Attribut 11 : statutory-holidays

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "nombre de jours fériés" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
statutory-holidays	numeric	9.0	15.0	11.0	10.0	18.0	11.1238...	12.0	11.0	Symetric

Figure 1.34: Attribut 11 : statutory-holidays du Data Set [labor.arff] .

BarChart

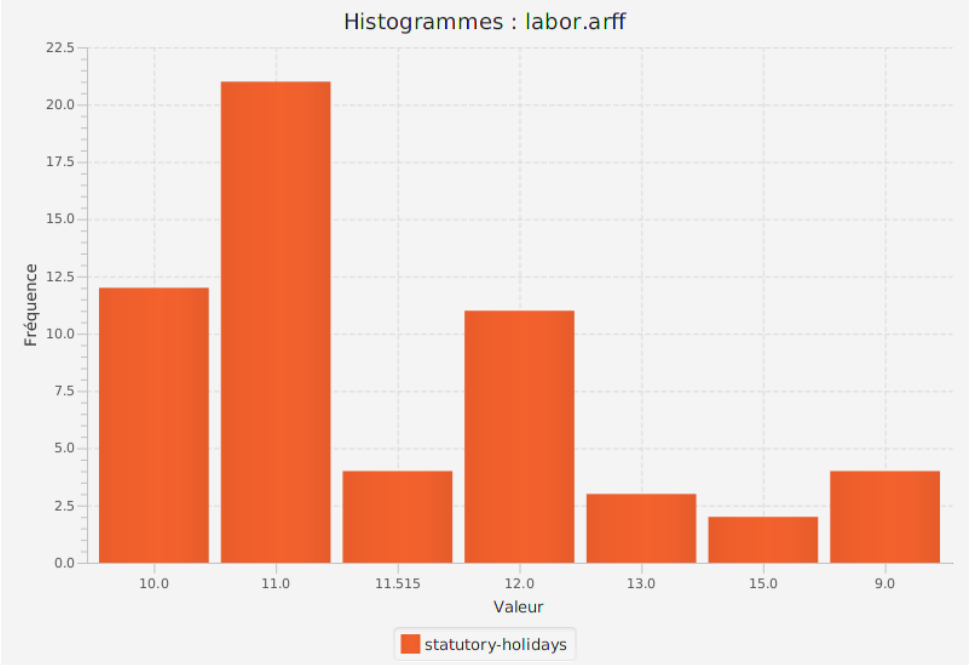

Figure 1.35: Histogramme de statutory-holidays du Data Set [labor.arff] .

discussion

• L'attribut statutory holidays est representatif des nombres en jours de vacances permises, on remarque que il varie entre min=10 et max=15 , d'après la répartition des données sur l'histogramme on peut interpréter l'existence d'un seul mode qui est : 11 ce qui implique attribut $\Rightarrow$ (uni-modal) ainsi que 10 et 11 ,13 couvre plus de la moitié (prés de 77%) de toutes la distribution des données.

• Maintenant à partir de la boite à moustache on peut facilement noter que la médiane et la moyenne sont égales , leur valeur étant 11 qui est aussi la valeur du mode cella nous permet de conclure que la distribution des données selon l'attribut "statutory holidays " est symétrique . (uniforme)

1.6.12 Attribut 12 : vacation

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "nombre de jours de congés payés" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
vacation	nominal	none	none	none	none	none	none	none	generous	none

Figure 1.36: Attribut 12 : vacation du Data Set [labor.arff] .

BarChart

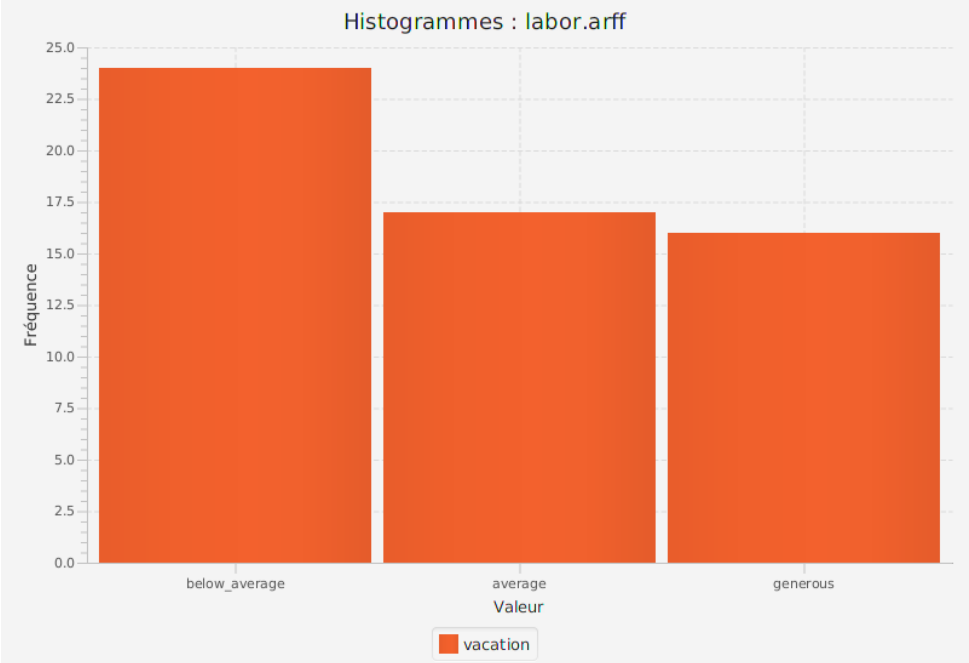

Figure 1.37: Histogramme de vacation du Data Set [labor.arff] .

discussion

• L'attribut vacation a trois valeurs possible **below-average,average,generous** , la distribution est assez uniforme tel que **below average** est la plus dominante avec 42% du tout le data set labor , suivie de prés de **average** avec 30% et **below-average** avec 28% .

• Vacation reste quand même un attribut uni-modal avec la valeur ("below average ") comme plus fréquente.

1.6.13 Attribut 13 : longterm-disability-assistance

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "employer's help during employee longterm disability " dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
longterm-disability-assistance	nominal	none	none	none	none	none	none	none	yes	none

Figure 1.38: Attribut 13 : longterm-disability-assistance du Data Set [labor.arff] .

La comparaison les valeurs de *[Mode , Médiane , Moyenne]*, n'est pas possible vu qu'il s'agit d'un attribut nominal.

BarChart

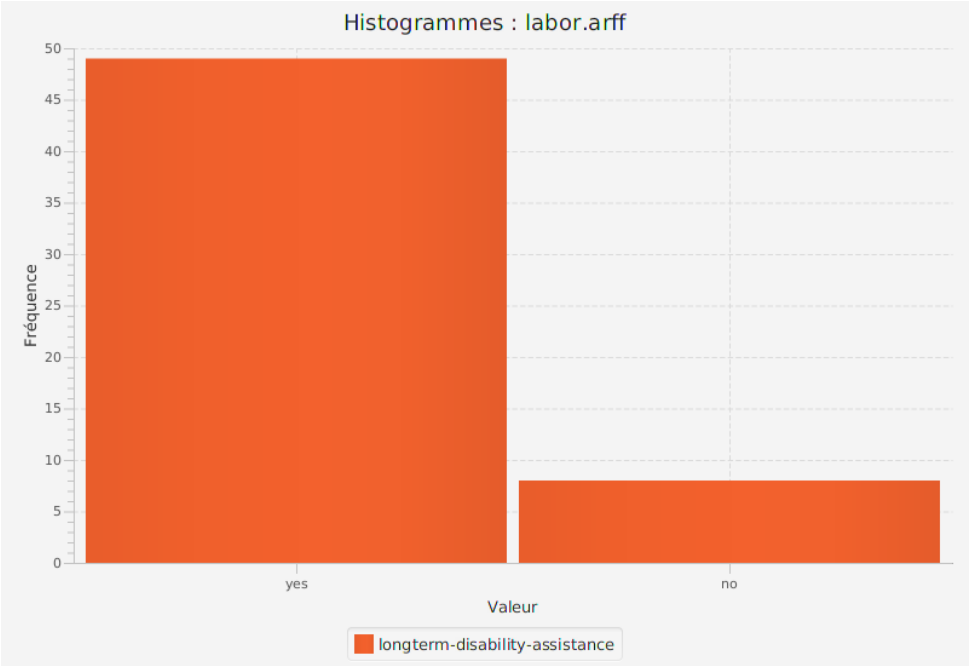

Figure 1.39: Histogramme de longterm-disability-assistance du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 2 valeurs possibles *("yes" , "no")*. la répartition penche nettement vers le "yes", c'est à dire une seule valeur domine tout le data set .
• D'ou l'on déduit que cet attribut est **Unimodale**, tel que mode = *("yes")*.

1.6.14 Attribut 14 : contribution-to-dental-plan

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "Contribution au régime de soins dentaires" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
contribution-to-dental-plan	nominal	none	none	none	none	none	none	none	half	none

Figure 1.40: Attribut 14 : contribution-to-dental-plan du Data Set [labor.arff] .

La comparaison les valeurs de *[Mode , Médiane , Moyenne]*, n'est pas possible vu qu'il s'agit d'un attribut nominal.

BarChart

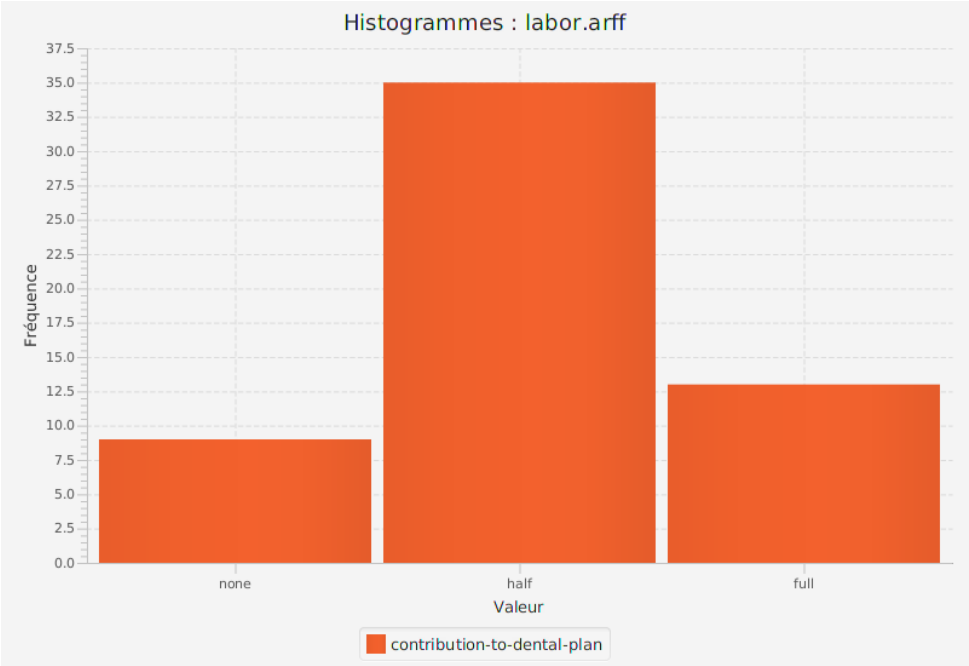

Figure 1.41: Histogramme de contribution-to-dental-plan du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 3 valeurs possibles *("none" , "half" , "full")*. la répartition n'est pas uniforme, c'est à dire il y a une différence significative entre le nombre d'instances des les valeurs possibles.
par exemple : la fréquence d'apparition de la valeur *("none")* comparé à celle de la valeur *("half")*.
• Il est aussi à noter que cet attribut est **Unimodale**, tel que mode = *("half")*.

1.6.15 Attribut 15 : bereavement-assistance

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "Contribution financière de l'employeur aux coûts du deuil" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
bereavement-assistance	nominal	none	none	none	none	none	none	none	yes	none

Figure 1.42: Attribut 15 : bereavement-assistance du Data Set [labor.arff] .

La comparaison les valeurs de *[Mode , Médiane , Moyenne]*, n'est pas possible vu qu'il s'agit d'un attribut nominal.

BarChart

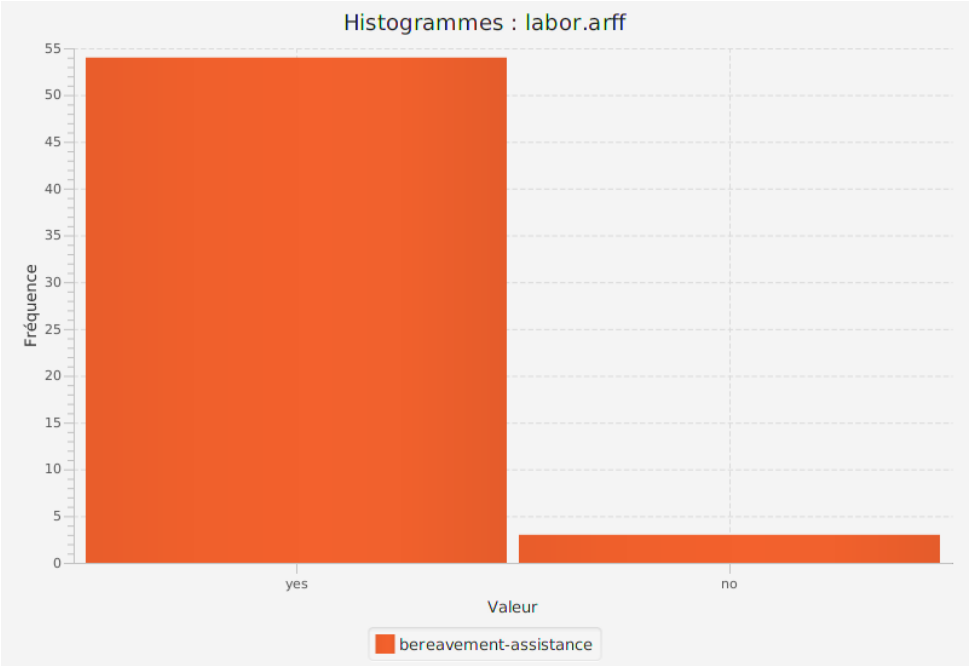

Figure 1.43: Histogramme de bereavement-assistance du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 2 valeurs possibles *("yes" , "no")*. la répartition n'est pas uniforme, c'est à dire les deux valeurs possibles n'apparaissent pas équitablement, autrement dis il y a un grand écart entre la fréquence de ces deux valeurs.
• Il est aussi à noter que cet attribut est **Unimodale**, tel que mode = *("yes")*.

1.6.16 Attribut 16 : contribution-to-health-plan

Caractéristiques

Nous résumons les caractéristiques de cet attribut : "contribution de l'employeur au régime de santé" dans le tableau ci-dessous:

Attribut	Type	Min	Max	Median	Q1	Q3	Mean	MidRan...	Mode	Symetric?
contribution-to-health-plan	nominal	none	none	none	none	none	none	none	full	none

Figure 1.44: Attribut 16 : contribution-to-health-plan du Data Set [labor.arff] .

La comparaison les valeurs de *[Mode , Médiane , Moyenne]*, n'est pas possible vu qu'il s'agit d'un attribut nominal.

BarChart

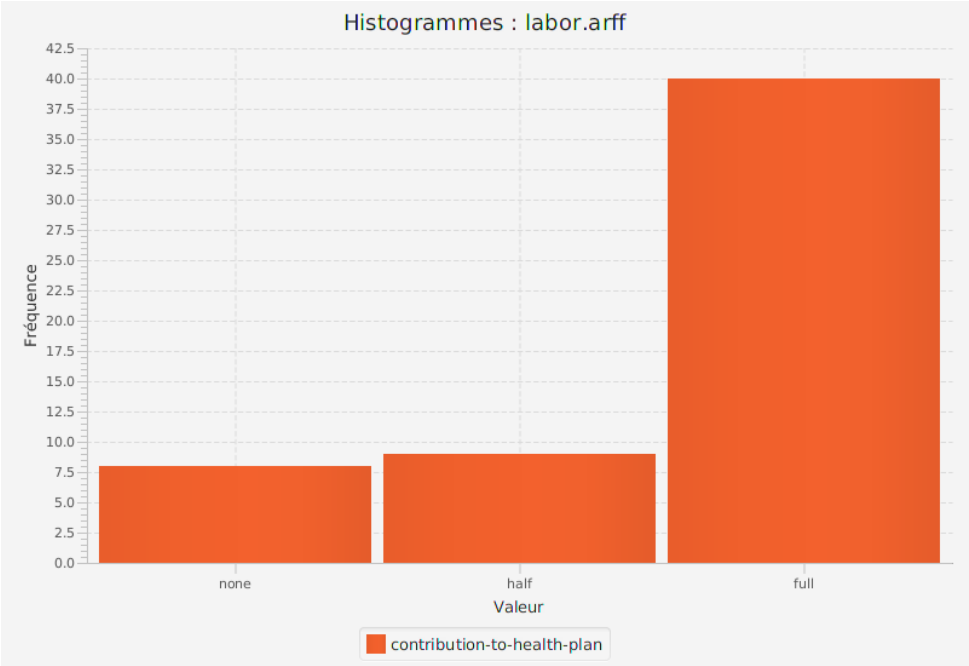

Figure 1.45: Histogramme de contribution-to-health-plan du Data Set [labor.arff] .

discussion

• L'on remarque que cet attribut peut prendre 3 valeurs possibles *("none" , "half" , "full")*. la répartition n'est pas uniforme, c'est à dire il y'a une grande différence entre le nombre d'instances des les valeurs possibles.
par exemple : la fréquence d'apparition de la valeur *("none")* comparé à celle de la valeur *("full")*.
• Il est aussi à noter que cet attribut est **Unimodale**, tel que mode = *("full")*.

1.6.17 La Class

Caractéristiques

Il s'agit de la classe, qui est dans la grande majorité des cas de type nominal, car très utilisé dans la classification et clusterisation.

BarChart

• L'on remarque que la class peut prendre 2 valeurs possibles *("bad" , "good")*. la répartition n'est pas très uniforme, c'est à dire qu'il y a plus d'instance labellisé *"good"* que celles labellisé *"bad"*.

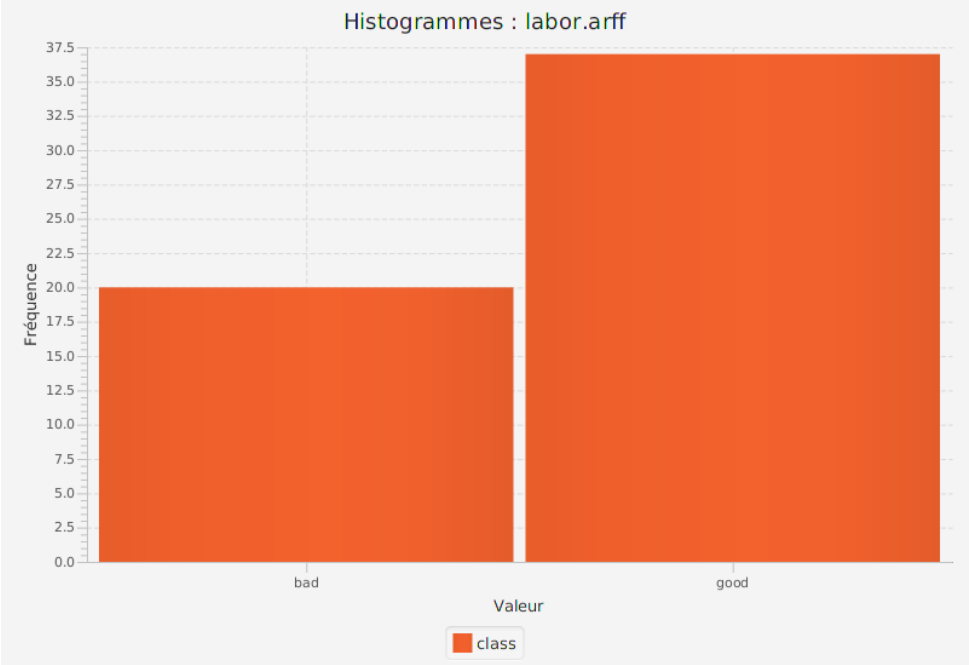

Figure 1.46: Histogramme de class du Data Set [labor.arff] .

1.7 Conclusion générale

Suite à l'analyse de données qu'on a pu effectuer sur le Data Set **labor.arff**, on se rend compte de la difficulté de la première étape du Data-Mining qui est la connaissance le pré-traitement de nos données , la connaissance de nos celles-ci a été bénéfique afin d'avoir une idée sur la distribution de ces derniers selon les attributs : par exemple on remarque que la plupart des instances pour la plupart des attributs ont des valeurs dominantes (mode) qui dans quelque cas couvre plus de la moitié ou encore le tiers des données .

Le Data Set **labor** est assez riche concernant les types d'attributs ,on a la moitié des attributs qui sont de type nominale (vacation,education-allowance ...) et l'autre partie est numérique (valeurs discrètes) (working-hours,duration..).

On note bien qu'il existe quatre attributs symétriques (duration,wage-increase-second-year,wage-increase-third-year,statutory-holidays) autrement dit il garde une distribution uniforme à gauche et à droite de la médiane et la moyenne , par contre trois sont négativement symétriques ce qui veut dire que la distribution se centre à gauche de la médiane (shift-differential,standby-pay,working-hours) et le seul attribut ayant une distribution symétrique positive est le (wage-increase-first-year) .

Aussi nous avons remarqué la présence des out-liers sur trois attributs (working-hours,shift-diffirentiel,wage increase) qui se traduit par des valeurs aberrantes se trouvant loin de la répartition de la distribution de nos données.

Bibliographie

[1] http://weka.sourceforge.net/doc.dev/
[2] http://www.info.univ-angers.fr/ gh/wstat/discr.php