République Algérienne Démocratique et Populaire

Ministère de l’Enseignement Supérieur et de la Recherche Scientifique

**Université des Sciences et de la Technologie Houari Boumediene**

Département Informatique

# Module : Data mining

**Rapport du projet-phase01-**

**Réalisation par :**

**-CHABANE Nouar 201400007379**

**-BEZGALI Meriem 201300005476**

**Groupe : 01**

1. **Introduction** :

L’extraction de données dépend étroitement du prétraitement des données, il est impératif de bien connaître nos données et les comprendre afin de pouvoir identifier par exemple les données manquantes, la répartition des différentes instances sur les valeurs de la médiane, le 1er quartile, le 3eme quartile, ainsi qu’une visualisation sous forme graphique de leurs distributions, c’est alors qu’intervient le prétraitement des données et l’analyse des données. Elles permettent d’obtenir une vision globale de l’ensemble de données, pour en suite effectuer l’étape du Datamining. Durant ce compte rendu, nous allons présenter une analyse d’un Dataset en faisant la conception d’une IHM qui permet d’afficher et d’effectuer cette analyse. L’analyse est faite selon 2 étapes :

* définir ce que les attributs représentent, identifier leurs types et la nature de leurs valeurs,
* utiliser des méthodes statistiques qui aboutissent par la génération de graphes et de calculer des paramètres de tendance centrale, qui sont la moyenne, la médiane et le mode qui donnent l’ordre de grandeur de l’ensemble des mesures, ainsi que les paramètres de dispersion, qui comportent le max, le min, les quartiles et d’autres.

1. **Identification des types des attributs du dataset et leurs valeurs :**

Le Dataset choisit est nommé « Labor.arff » utilisé pour apprendre si la description d’un contrat acceptable ou pas, nous l’avons choisit car il possède des attributs de différent types ainsi que des valeurs manquantes, il est constitué de :

* 57 instances.
* 17 attributs (pas d’attributs manquants).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Attribut | Type | Description |
| 0 | 'duration' | Numeric | duration of agreement [1..7] |
| 1 | 'wage-increase-first-year' | Numeric | wage increase in first year of contract [2.0..7.0] |
| 2 | 'wage-increase-second-year' | Numeric | wage increase in second year of contract [2.0..7.0] |
| 3 | 'wage-increase-third-year' | Numeric | wage increase in third year of contract [2.0..7.0] |
| 4 | 'cost-of-living-adjustment' | Nominal | cost of living allowance [none, tcf, tc] |
| 5 | working-hours' | Numeric | number of working hours during week [35..40] |
| 6 | ‘Pension’ | Nominal | employer contributions to pension plan [none, ret\_allw, empl\_contr] |
| 7 | 'standby-pay | Numeric | standby pay [2..25] |
| 8 | ‘shift-differential' | Numeric | Shift differencial : supplement for work on II and III shift [1..25] |
| 9 | 'education-allowance' | Nominal | education allowance [true false] |
| 10 | ‘statutory-holidays ‘ | Numeric | number of statutory holidays  [9..15] |
| 11 | 'vacation' | Nominal | number of paid vacation days  [ba, avg, gnr] |
| 12 | 'longterm-disability-assistance' | Nominal | employer's help during employee longterm disability [true, false] |
| 13 | 'contribution-to-dental-plan’ | Nominal | employers contribution towards the dental plannone, [half, full] |
| 14 | 'bereavement-assistance' | Nominal | employer's financial contribution towards the  covering the costs of bereavement  [true, false] |
| 15 | 'contribution-to-health-plan' | Nominal | employer's contribution towards the health plan [none, half, full] |
| 16 | ‘class’ | Nominal | Contract’s classification [bad,good] |

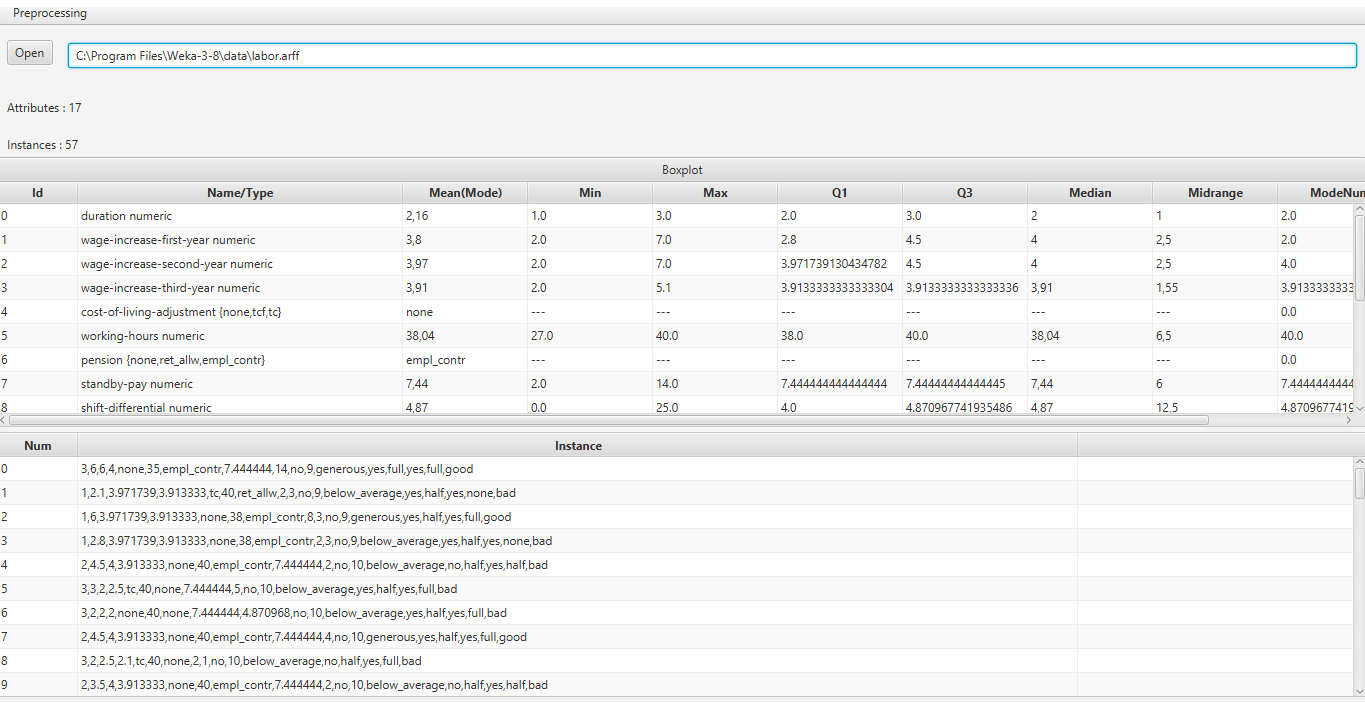
1. **Analyse statistiques et calcule des paramètres de tendance centrale :**

Dans ce qui suit nous allons visualiser les résultats statistiques obtenus dans l’IHM réalisée. L’application permet de lire le DataSet, récupérer les instances et les afficher ainsi que les attributs et leurs types.

* 1. **Lecture et affichage des Datasets**

Pour afficher le dataset voulu nous disposons d’un bouton « Open » qui permet de parcourir les dataset ayant l’extension « \*.arff »

 Figure1.Bouton « Open »

 Figure2.Interface de l’application

* 1. **Prétraitement**

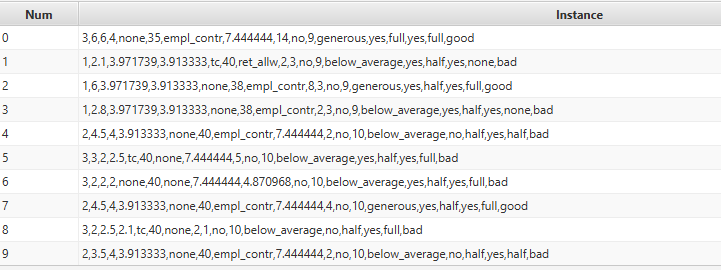
Dans cette partie, il s’agit de normaliser les attributs numériques et remplacer les valeurs manquantes.

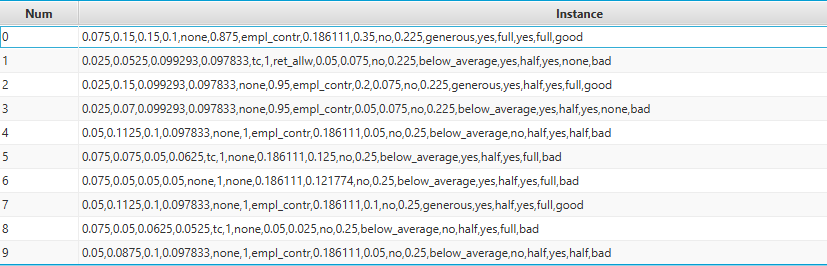
1. Remplacement des valeurs manquantes : Les valeurs manquantes dans les Datasets sont désignées par le symbole « ? ». Il s’agit de remplacer chaque valeur par la moyenne de l’attribut des instances appartenant à la même classe quand il s’agit d’un attribut numérique et par la valeur la plus fréquente dans la même classe pour les attributs nominaux.



 Figure3.remplacement des valeurs manquantes

1. La normalisation : Permet d’ajuster une série de valeurs suivant une fonction de transformation pour les rendre sur un intervalle entre 0 et 1. On choisit d’appliquer la formule (valeur-min) / (max-min). Ce traitement est appliqué aux attributs numériques.

****

****Figure4.Normalisation du dataset

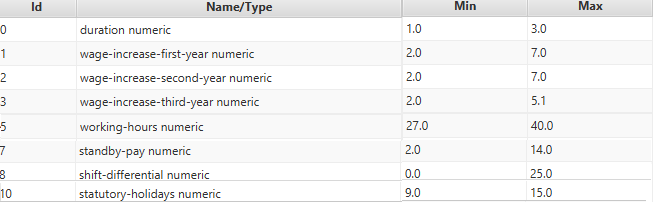
* 1. **Tendances centrales**

Afin d’analyse le dataset « Labor » nous allons présenter les valeurs des tendances centrales et essayer de savoir si certain attributs sont symétrique ou asymétrique (gauche ou droite).

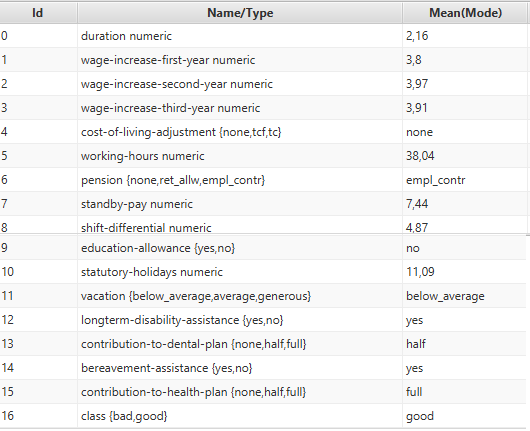
1. **Max et Min**

* **Max =** La plus grande valeur que peut prendre un attribut**.**
* **Min** = La plus petite valeur que peut prendre un attribut.

Dans le cas du dataset labor les min et max pour chaque attribut numérique sont :

Figure5.Max et Min des attributs numériques

1. **Moyenne**
2. **Cas de valeurs numériques** : La moyenne est calculée en divisant la somme de toutes les valeurs d’un attribut par le nombre total de valeurs que peut prendre ce dernier. 𝑥̅=(𝑥1+𝑥2+⋯+𝑥𝑁)/N
3. **Cas de valeurs nominales** : On affiche le mode qui est la valeur qui a le plus d’occurrence pour chaque attribut.

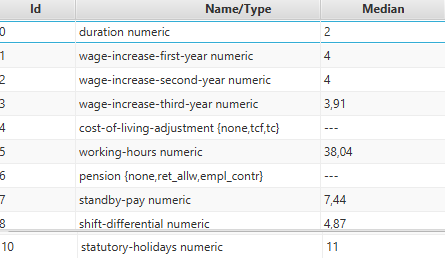
 Figure6.Moyenne (pour valeurs numérique) et mode (valeurs nominales)

1. **Médiane**

La médiane partage un ensemble de valeurs numériques rangées par ordre croissant en deux sous-ensembles de taille égale et est aussi appelée Q2 (le 2eme quartile).

Soit N la taille de l’ensemble des valeurs et M la médiane :

* 𝑆𝑖 𝑁 = 2𝑝 𝑎𝑙𝑜𝑟𝑠 𝑀 (𝑝+1) eme 𝑣𝑎𝑙eur.
* 𝑆𝑖 𝑁 = 2𝑝+1 𝑎𝑙𝑜𝑟𝑠 𝑀= 𝑙𝑎 𝑚𝑜𝑦𝑒𝑛𝑛𝑒 𝑒𝑛𝑡𝑟𝑒 𝑙𝑎 𝑝eme 𝑒𝑡 𝑙𝑎 (𝑝+1) eme 𝑣𝑎𝑙𝑒𝑢𝑟.

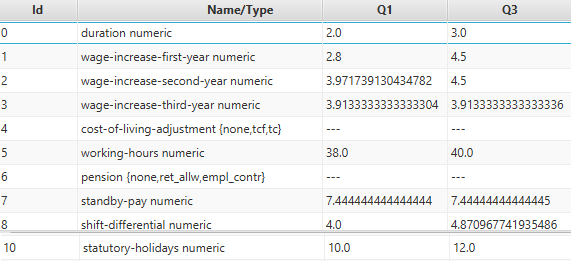
**** Figure7.Mediane des attributs numériques du dataset

1. **Quartiles (Q1 et Q3)**

Les quartiles sont les valeurs qui divisent un ensemble de valeurs en 4 sous-ensembles de même taille. Pour calculer les quartiles, il faut d’abord que les valeurs soient rangées par ordre croissant.

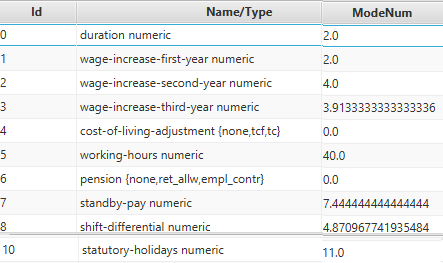
Soit N la taille de l’ensemble des valeurs de type numérique. On distingue 4 cas :

* 𝑆𝑖 𝑁=4𝑝 𝑎𝑙𝑜𝑟𝑠 {𝑄1= 𝑚𝑜𝑦𝑒𝑛𝑛𝑒 𝑒𝑛𝑡𝑟𝑒 𝑙𝑎 𝑝𝑒 𝑒𝑡 (𝑝+1)eme 𝑣𝑎𝑙𝑒𝑢𝑟 𝑄3=𝑚𝑜𝑦𝑒𝑛𝑛𝑒 𝑒𝑛𝑡𝑟𝑒 𝑙𝑎 (3𝑝)𝑒me 𝑣𝑎𝑙𝑒𝑢𝑟 𝑒𝑡 𝑙𝑎 (3𝑝+1)𝑒me 𝑣𝑎𝑙𝑒𝑢𝑟 }
* 𝑆𝑖 𝑁=4𝑝+1 𝑎𝑙𝑜𝑟𝑠 {𝑄1= (𝑝+1)eme 𝑣𝑎𝑙𝑒𝑢𝑟 𝑄3= (3𝑝+1)𝑒me 𝑣𝑎𝑙𝑒𝑢𝑟 }
* 𝑆𝑖 𝑁=4𝑝+2 𝑎𝑙𝑜𝑟𝑠 {𝑄1= (𝑝+1)eme 𝑣𝑎𝑙𝑒𝑢𝑟 𝑄3= (3𝑝+2)𝑒me 𝑣𝑎𝑙𝑒𝑢𝑟 }
* 𝑆𝑖 𝑁=4𝑝+3 𝑎𝑙𝑜𝑟𝑠 {𝑄1= (𝑝+1)eme 𝑣𝑎𝑙𝑒𝑢𝑟 𝑄3= (3𝑝+3)𝑒me 𝑣𝑎𝑙𝑒𝑢𝑟}

**** Figure8.les Q1 et Q3 des attributs numériques

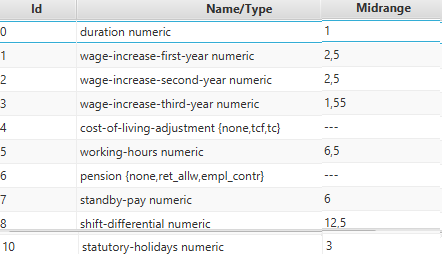
1. **Mode**

Le mode représente la valeur la plus fréquente quel que soit le type de cette valeur numérique ou nominale (le cas nominal est déjà traité précédemment).

**** Figure8.le mode pour chaque attribut numérique

1. **Midrange**

C’est la moyenne arithmétique des valeurs les plus grandes et les plus petites pour chaque attribut numérique, et est calculé comme suit : Midrange = (max-min)/2 , et on prend la valeur qui se trouve a cette position après avoir effectuer un trie des valeurs numériques.

 Figure9.Le midrange pour chaque attribut numérique

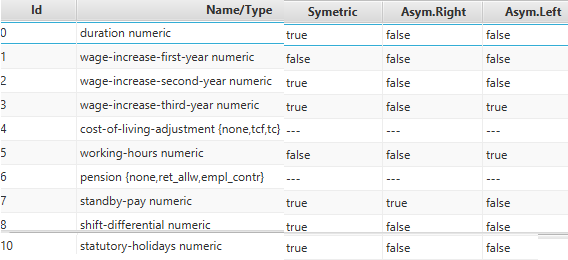
* 1. **Analyse et Graphes**

Une distribution d’un ensemble de valeur est dite symétrique si les valeurs observées se répartissent de façon uniforme autour des trois valeurs centrales : la moyenne, le mode et la médiane.

On distingue trois types de distributions selon qu’elles sont asymétriques (à droite, à gauche) ou symétriques comme le montrent les formules suivante :

* 𝑀𝑜𝑦𝑒𝑛𝑛𝑒 = 𝑀é𝑑𝑖𝑎𝑛𝑒 = 𝑚𝑜𝑑𝑒 (veut dire 𝑆𝑦𝑚é𝑡𝑟𝑖𝑞𝑢𝑒)
* 𝑀𝑜𝑦𝑒𝑛𝑛𝑒 < 𝑀é𝑑𝑖𝑎𝑛𝑒 < 𝑀𝑜𝑑𝑒 (veut dire 𝐴𝑠𝑦𝑚é𝑡𝑟𝑖𝑞𝑢𝑒 à 𝑔𝑎𝑢𝑐ℎ𝑒)
* 𝑀𝑜𝑦𝑒𝑛𝑛𝑒 > 𝑀é𝑑𝑖𝑎𝑛𝑒 > 𝑀𝑜𝑑 (veut dire 𝐴𝑠𝑦𝑚𝑒𝑡𝑟𝑖𝑞𝑢𝑒 à 𝑑𝑟𝑜𝑖𝑡𝑒)

Pour chaque cas cité précédemment nous avons implémenté une fonction qui vérifie la symétrie et l’asymétrie à gauche et à droite d’un attribut numérique, pour le cas de la symétrie on tolère 0.25 comme marge d’erreur, nous avons obtenu les résultats suivant pour chaque attribut numériques.

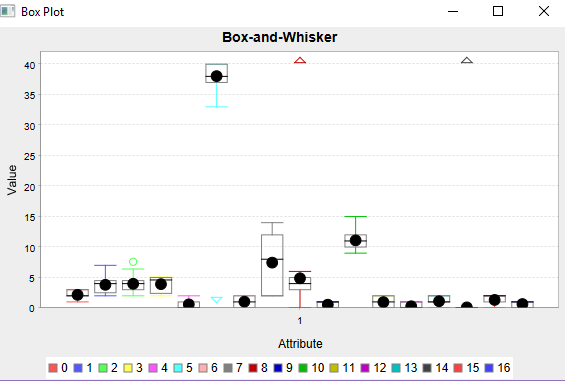
Figure10.La symétrie et asymétrie des attributs numérique du dataset

Remarque : Pour une marge d’erreur = 0.25 nous avons obtenus un total de 6/8 attributs numérique qui sont symétriques.

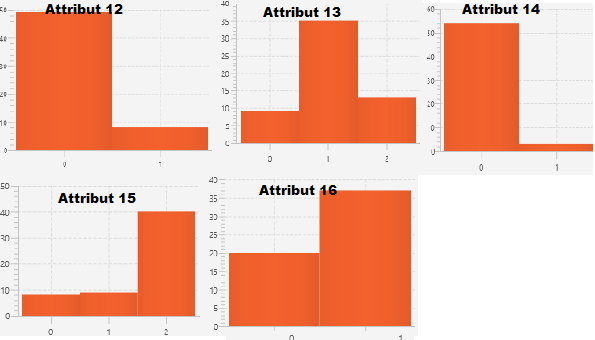
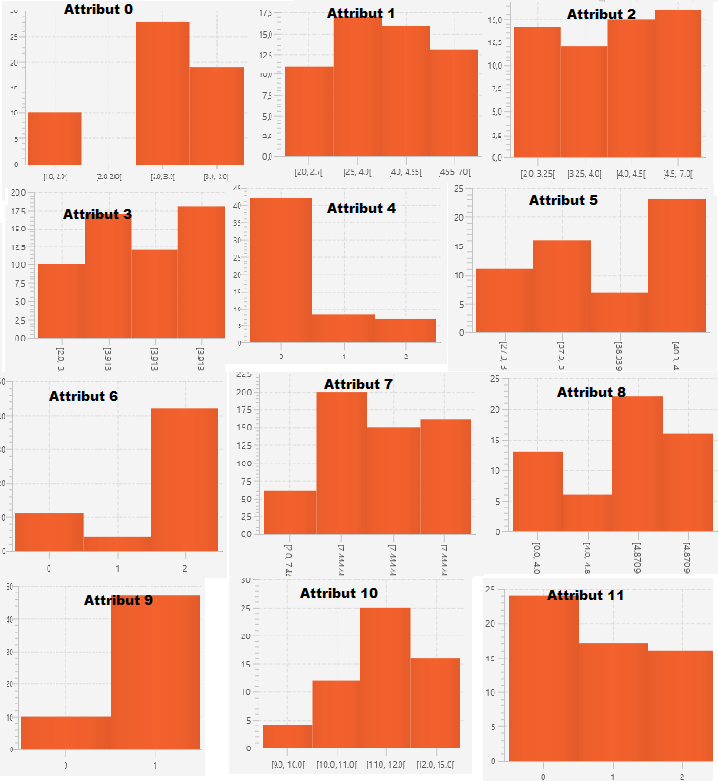
1. **Boit à moustaches**

La boite à moustache est utile lorsqu’on veut visualiser la symétrie, la dispersion, les valeurs aberrantes, et les quartiles. Le trait de chaque boite représente la médiane. Les ronds et triangles se trouvant hors de la boite représentent les valeurs aberrantes.

Les attributs d’identificateurs « 2 » « 8 » et « 14 » contiennent des valeurs aberrantes. Concernant la symétrie, on remarque que le trait de médiane est presque au milieu pour les attributs d’indice « 0 » « 2 » « 3 » « 7 » « 8 » « 10 », ce qui confirme le résultat trouvé précédemment dans les calcule de symétrie avec les fonctions implémentés.

**** Figure11.Boxplot du dataset « labor »

1. **Les histogrammes**

****Figure12. Les histogrammes des 17 attributs de dataset « labor »

1. **Cas valeurs numérique :** Nous avons effectué une discrétisation afin de dessiner les graphes des attributs numérique en utilisant la médiane à chaque division de classe, et nous l’avons applique 3 fois ce qui donne en résultats 4 classes donc 4 barres au total pour chaque attribut.
2. **Cas valeurs nominales :** On considère par exemple le cas de l’attribut « Class » d’indice « 16» qui peut prendre 2 valeurs soit « bad» soit « good », la première valeurs et codifié en 0 la deuxième en 1, en générale autant qu’il y’ a de valeur pour l’attribut on leurs affecte un chiffre suivant l’ordre croissant.

On constante qu’on peut confirmer les résultats obtenu lors du calcule du mode précédemment en visualisant les graphes pour chaque attribut pour voir les valeurs les plus fréquente, aussi on remarque que les valeurs prise en générale par les attributs sont vraiment petite et proche l’une de l’autre ce qui montre que les valeurs sont répartie sur une plage de nombre assez petite.

Pour le cas de la symétrie ou asymétrie veut qu’on a appliqué une discrétisation on ne peut pas la remarqué sur les graphes, mais on la détecte avec les fonctions implémenté précédemment.

1. **Conclusion**

Nous avons, lors de ce TP pu atteindre nos objectifs fixés au départ :

* Extraire et affichage du contenu et la description d’un Dataset.
* Calculer les mesures de tendance centrale et déduire la symétrie ou asymétrie des attributs numériques.
* Affichage des graphes (boite à moustache et histogramme) pour chaque attribut et déduire les valeurs les plus fréquentes ainsi que les quartiles et les médianes et valeurs aberrantes.
* Raffiner le dataset pour effectuer l’étape du mining et obtenir de bons résultats.

Cette phase du TP nous a permis de faire une analyse sur les datasets « labor » et apprendre de manière générale comment effectuer analyse sur n’importe quel dataset donné, on a pu aussi voir les caractéristiques qui peuvent être contenue dans les attributs d’un dataset en générale, et de traiter les valeurs manquantes et d’effectuer une normalisation dans l’étape du prétraitement, on a également pris conscience de l’importance de l’analyse et du prétraitement avant de faire du Data Mining, afin de ne familiarisé avec le dataset et essayer d’extraire le modèle le plus précis possible.