Université des Sciences et de la Technologie HOUARI BOUMEDIENE
Faculté d'Electronique et d'Informatique
Département d'Informatique

Data Mining

Etude Exploratoire d'un Dataset

Partie N° : 1

Etude Exploratoire d'un Dataset

- ## Langage de programmation & environnement

Recommandations

JDK 11

https://www.oracle.com/java/technologies/javase/jdk11-archive-downloads.html

Netbeans 12.5

https://netbeans.apache.org/download/nb125/nb125.html

Jfreechart 1.5.2

http://www.jfree.org/jfreechart/jfreechart-demo-1.5.2-jar-with-dependencies.jar

Etude Exploratoire d'un Dataset

- **Dataset à étudier :** seeds

N° instance	**1er** attribut	**2ème** attribut	**3ème** attribut	**4ème** attribut	**5ème** attribut	**6ème** attribut	**7ème** attribut	**Classe**
1	15.26	14.84	0.871	5.763	3.312	2.221	5.22	Kama
2	14.88	14.57	0.8811	5.554	3.333	1.018	4.956	Kama
3	14.29	14.09	0.905	5.291	3.337	2.699	4.825	Kama
...	...	...	...	...	...	...	...	...
70	12.73	13.75	0.8458	5.412	2.882	3.533	5.067	Kama
71	17.63	15.98	0.8673	6.191	3.561	4.076	6.06	Rosa
72	16.84	15.67	0.8623	5.998	3.484	4.675	5.877	Rosa
...	...	...	...	...	...	...	...	...
140	16.23	15.18	0.885	5.872	3.472	3.769	5.922	Rosa
141	13.07	13.92	0.848	5.472	2.994	5.304	5.395	Canadian
142	13.32	13.94	0.8613	5.541	3.073	7.035	5.44	Canadian
...	...	...	...	...	...	...	...	...
209	11.84	13.21	0.8521	5.175	2.836	3.598	5.044	Canadian
210	12.3	13.34	0.8684	5.243	2.974	5.637	5.063	Canadian

Etude Exploratoire d'un Dataset

Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Manipuler le dataset et visualiser son contenu.

Ouverture du dataset

Les extensions possibles .csv & .txt
A partir d'un disque ou d'une URL valide

Affichage du dataset

Indiquer les numéros des instances, les attributs & la classe

Ajout, Modification, Suppression d'une instance

Sauvegarde du dataset

Etude Exploratoire d'un Dataset

Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Fournir une description du dataset et de chacun de ses attributs.

Inclure dans le rapport

Description du dataset

Informations générales sur le dataset
Nombre d'instances
Nombre d'attributs
Nombre de classes
Distribution des classes (nombre et pourcentage d'instances par classe)
Valeurs manquantes

Description des attributs

Numéro, nom, description, type et valeurs possibles de chaque attribut
Types possibles (Nominal, Binaire symétrique, Binaire asymétrique, Numérique) – (Qualitatif, Quantitatif) – (Discret, continu)

Etude Exploratoire d'un Dataset

Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Pour chaque attribut, calculer les mesures de tendance centrale et en déduire les symétries.

Inclure dans le rapport

Moyenne (mean)

Moyenne tronquée (trimmed mean)
La moyenne tronquée est la moyenne obtenue après avoir éliminé les valeurs extrêmes. Une troncature à 2% signifie qu'on ignore 2% des données les plus éloignées.

Médiane (median)

Mode (mode)
Unimodale – multimodale (bimodale, trimodale, …)

Milieu de l'étendue (midrange)
La moyenne des deux valeurs extrêmes

Etude Exploratoire d'un Dataset

- ## Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Pour chaque attribut, calculer les mesures de tendance centrale et en déduire les symétries.

Inclure dans le rapport

Données symétriques (symmetric data)

Données asymétriques à droite (positively skewed data)

Données asymétriques à gauche (negatively skewed data)

Etude Exploratoire d'un Dataset

Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Pour chaque attribut, calculer les mesures de dispersion et en déduire les données aberrantes (outliers).

Inclure dans le rapport

- **Etendue** (range)
- **Quartiles** (five-number summary)
  $Q_0,\ Q_1,\ Q_2,\ Q_3,\ Q_4$
- **Ecart interquartile** (interquartile range IQR)
- **Variance**
- **Ecart-type**

Etude Exploratoire d'un Dataset

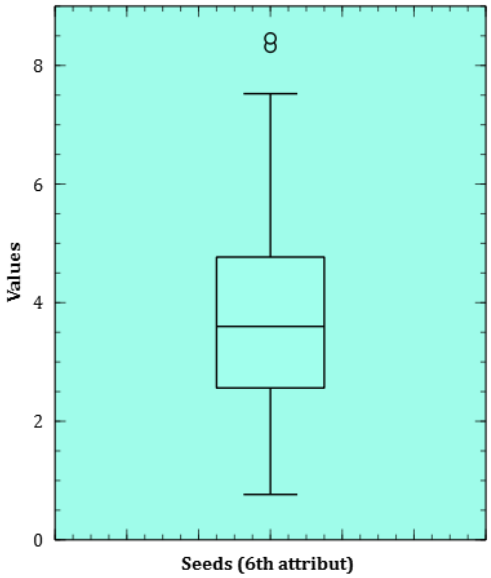
Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Pour chaque attribut, construire une boîte à moustache et afficher les données aberrantes.

Boîte à moustache (boxplot)

Inclure toutes les boîtes à moustaches dans le rapport



Etude Exploratoire d'un Dataset

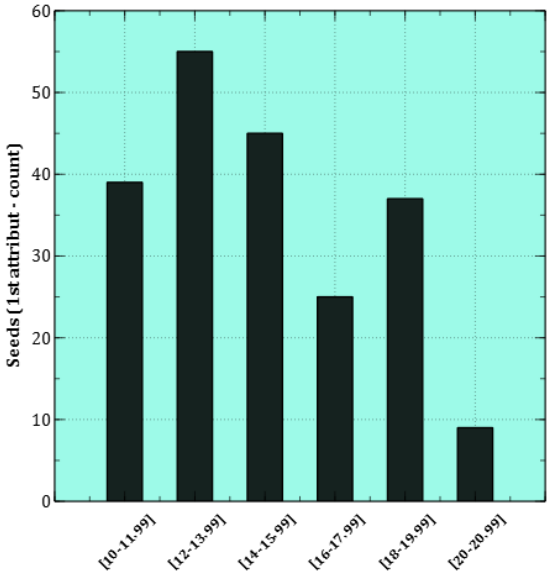
- ## Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Pour chaque attribut, construire un histogramme et visualiser la distribution des données

Histogramme (histogram)

Inclure tous les histogrammes (+interprétation de chaque histogramme) dans le rapport



Etude Exploratoire d'un Dataset

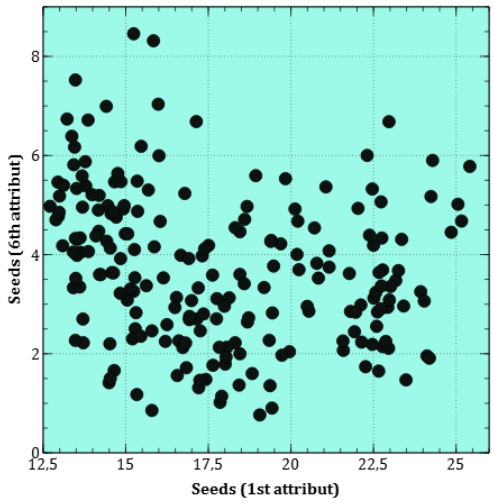
Travail à réaliser

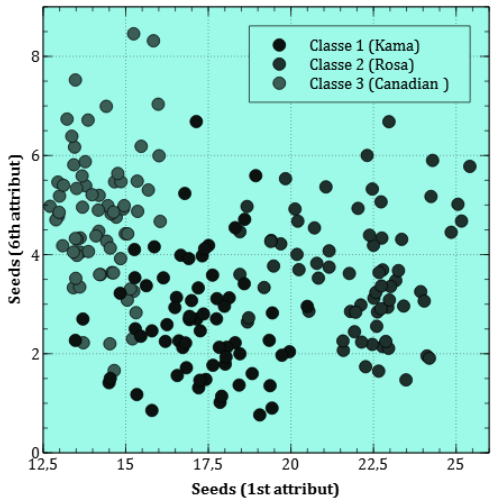
Concevoir et programmer en Java une IHM permettant de :

Construire et afficher des diagrammes de dispersion des données et en déduire les corrélations entre les attributs.

Diagramme de dispersion (scatter plot)

Inclure tous les diagrammes de dispersion dans le rapport





Etude Exploratoire d'un Dataset

- ## Travail à réaliser

Concevoir et programmer en Java une IHM permettant de :

Construire et afficher des diagrammes de dispersion des données et en déduire les corrélations entre les attributs.

Inclure dans le rapport

Analyse de corrélation

Coefficient de corrélation & covariance

Corrélation positive (forte, moyenne, faible), corrélation négative, pas de corrélation