Classification des données relatives aux passagers du Titanic et chances de survie à la catastrophe

John Sherlock, Manoj Muniswamaiah, Lauren Clarke, Shawn Cicoria
École Seidenberg d'informatique et de systèmes d'information Université Pace
White Plains, New York, États-Unis
{js20454w, mm42526w, lc18948w}@pace.edu, shawn@cicoria.com

Résumé *– Bien que le naufrage du Titanic ait eu lieu il y a un peu plus de 100 ans, il continue d'attirer les chercheurs qui cherchent à comprendre pourquoi certains passagers ont survécu tandis que d'autres ont péri. À l'aide d'outils modernes d'exploration de données (Weka) et d'un ensemble de données disponible, nous examinons quels facteurs ou classifications des passagers ont une relation convaincante avec la survie des passagers qui ont pris ce voyage fatidique le 10 avril 1912. L'analyse vise à identifier les caractéristiques des passagers (classe de cabine, âge et point de départ) et leur relation avec les chances de survie lors du naufrage.*

Mots-clés : exploration de données ; Titanic ; classification ; Kaggle ; Weka

I. INTRODUCTION

Le Titanic est un navire qui a fait naufrage lors de son voyage inaugural dans l'Atlantique Nord le 15 avril 1912, causant la mort de 1 502 des 2 224 passagers et membres d'équipage[2]. Bien qu'il existe des conclusions concernant la cause du naufrage, l'analyse des données sur les facteurs qui ont influé sur la survie des passagers se poursuit à ce jour[2,3]. L'approche adoptée consiste à utiliser un ensemble de données accessibles au public provenant d'un site web appelé Kaggle[4] et l'outil d'exploration de données Weka[5]. Nous nous sommes concentrés sur l'analyse par arbre de décision et l'analyse par grappes après examen et normalisation des données.

A. Kaggle – Modélisation prédictive et analyse

Kaggle propose aux entreprises et autres entités des services de crowdsourcing en matière d'exploration de données, d'apprentissage automatique et d'analyse. Il offre parfois des prix (par exemple, GE a offert un prix de 200 000 dollars via Kaggle dans le cadre d'un concours[1]).

B. Weka - Waikato Environment for Knowledge Analysis

L'outil Weka fournit un ensemble d'outils d'apprentissage automatique et d'exploration de données. Disponible gratuitement, il est basé sur Java, ce qui lui permet de fonctionner sur les plateformes qui prennent en charge Java. Il est principalement maintenu et pris en charge par des chercheurs de l'université de Waikato.

II. DONNÉES ET MÉTHODOLOGIE

A. Exemple de données provenant de Kaggle

Voici une représentation de l'ensemble de données de test fourni au format CSV (valeurs séparées par des virgules) par Kaggle et comprenant 891 lignes de données (un sous-ensemble de la liste complète des passagers). La structure du fichier avec des exemples de lignes est présentée dans les 3 tableaux suivants.

Tableau I. Échantillon de données Kaggle Titanic

passengerid	survived	pclass	nom	sex
1	0	3	Braund, M.	homme
2	1	1	Cummings, Mme	femme
3	1	3	Heikkinen, M	Femme

Tableau II. Échantillon de données Kaggle Titanic (suite)

Âge	frères et sœurs	parch	ticket
22	1	0	A/5 2117
38	1	0	PC 17599
26	0	0	STON/O2

Tableau III. Échantillon de données Kaggle Titanic (suite)

tarif	cabine	embarqué
7,25		S
71,2833	C85	C
7,925		S

B. Normalisation des données

L'ensemble de données a été modifié afin de créer des colonnes nominales à partir de certaines colonnes numériques, dans le but de faciliter leur utilisation dans Weka pour l'analyse arborescente et l'analyse simple par grappes.

La modification a été effectuée afin de faciliter l'utilisation dans Weka pour l'analyse arborescente et l'analyse simple par grappes. Le tableau suivant identifie les conversions et autres modifications.

TABLEAU I. ENSEMBLE DE DONNÉES KAGGLE TYPES DE DONNÉES NORMALISÉS

Champ	Modification	Commentaire
ID passager	Ignoré	Non nécessaire
Survivant	Converti en NON/OUI	Nécessaire nominal Identifiant
Pclass	Supprimé -> classe créée à la place	Colonne non nominale requise identifiant
Classe	Nouvelle colonne	Calcul simple basé sur « pclass »
Groupe d'âge	Basé sur une formule ; certains valeurs non fournies. Mais a abouti à 4 groupes autre que Inconnu (Enfant, Adolescent, Adulte, Personne âgée)	Les valeurs ont été attribuées de manière arbitraire ce qui suit : =IF(F2="", « Inconnu », SI(F2<10, « Enfant », SI(F2<20, « Adolescent », IF(F2<50, « Adulte », « Personne âgée »))))
Ecode	Supprimé -> classe créée Embarqué	Nominal nécessaire identifiant
Embarqué	Nouvelle colonne qui a converti Ecode en nom du point de départ pour le passager	

C. *Ensemble de données d'analyse normalisé*

Une fois converti, l'ensemble de données final utilisé pour l'analyse dans l'outil Weka est illustré ci-dessous, avec les premières lignes affichées.

TABLEAU II. EXEMPLE D'ENSEMBLE DE DONNÉES NORMALISÉES

Identifiant passager	Survivant	Pclass	Classe
1	Non	3	3e
2	Oui	1	1er
3	Oui	3	Erd

TABLEAU III. EXEMPLE D'ENSEMBLE DE DONNÉES NORMALISÉES (SUITE)

Sexe	Âge	Groupe d'âge	Code écologique	Embarqué
Masculin	22	adulte	S	Southhamptom
Femme	38	adulte	C	Cherbourg
Femme	26	adulte	S	Southhamtom

D. *Format de fichier ARFF Weka*

Le tableau est ensuite converti et enregistré au format de fichier Weka Attribute-Relation File Format (ARFF). Le fichier ARFF utilisé est présenté à l'annexe E. Les principales caractéristiques du format de fichier ARFF, qui facilitent l'exploration des données dans l'outil Weka, sont l'identification des types de données et, au sein de ces champs, l'ordre des valeurs nominales.

III. ANALYSE DES DONNÉES

A. *Classification par arbre de décision*

À l'aide de Weka, nous avons généré un arbre J48[6] (implémentation C.45) qui a donné lieu à la sortie du classificateur représentée dans l'annexe G

- Résultat du classificateur J48. Le diagramme de l'arbre J48 présenté dans la figure 2 ci-dessous illustre le chemin de classification suggéré par les données.

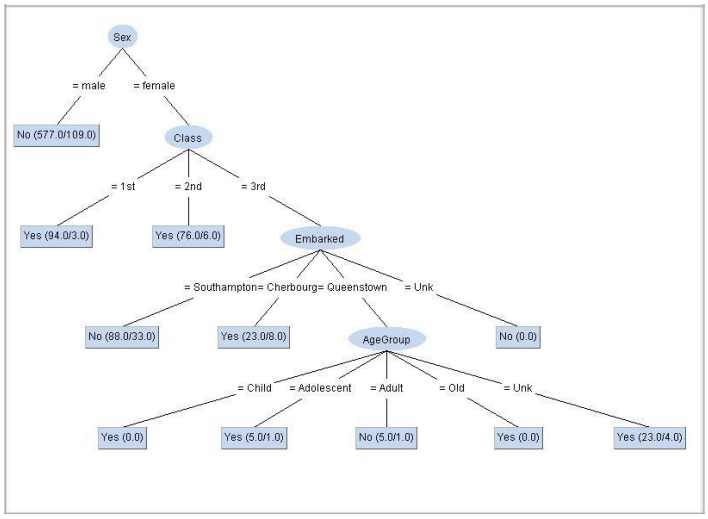


Fig. 1. Diagramme du classificateur J48

B. *Classificateur J48 - Conclusions initiales*

D'après les résultats de l'analyse J48, il était clair que l'association la plus significative en matière de survie était liée au sexe, le fait d'être une femme étant le classificateur le plus significatif. Nous avons ensuite examiné l'analyse par grappes afin de mettre en évidence d'autres relations.

C. *Analyse simple par grappes K-means*

Le regroupement des données en fonction des classifications et l'utilisation d'associations simples issues de l'analyse par grappes permettent de comprendre les données. Bien qu'une association puisse être forte grâce à cette analyse, il n'est pas possible de conclure à une véritable relation de cause à effet.

D. *Résultats de l'analyse simple K Means*

Pour notre analyse par grappes, nous avons choisi la méthode simple K Means, simplement pour des raisons de simplicité. Les résultats textuels de la méthode simple K Means sont inclus dans l'annexe H. Les visualisations sont également présentées dans les sections suivantes.

À l'aide du diagramme de regroupement, nous pouvons analyser visuellement les regroupements pour déterminer les relations au sein de l'ensemble de données. La force de la classification et du regroupement est indiquée visuellement ainsi que dans le résultat textuel. Cette relation de regroupement peut être utilisée pour conclure qu'il existe une certaine relation, mais pas une relation de cause à effet.

E. *Survivants vs sexe*

Visuellement, nous constatons de manière assez spectaculaire que le sexe des passagers présente un regroupement significatif autour des chances de survie. Cela a également été montré dans l'arbre J48. La figure 2 ci-dessous illustre le regroupement significatif entre le sexe et les chances de survie. Que cela soit prévisible ou non, il faudrait mener une analyse complémentaire en sciences sociales pour comprendre pourquoi un sexe peut mieux s'en sortir dans ces situations traumatisantes.

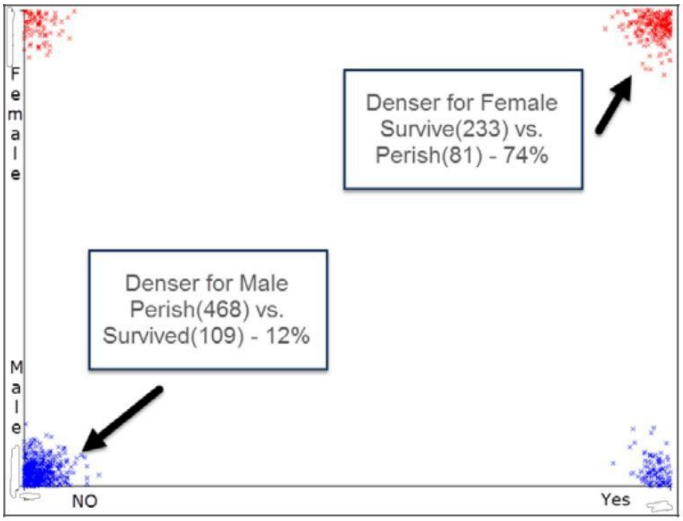


Fig. 2. Méthode K simple – Classification des survivants par sexe

F. Survivants vs classe

Sans surprise, la classe de cabine présentait un regroupement significatif, les cabines de niveau inférieur affichant un poids significatif en faveur de la non-survie. Cela est illustré dans la figure 3 ci-dessous, avec un regroupement assez dominant pour les personnes de 3e classe qui n'ont pas survécu. Et un regroupement assez clair pour les personnes de 1re classe qui ont survécu. Nous pouvons émettre des hypothèses sur ce résultat, peut-être en fonction de l'emplacement physique ou d'autres faits concernant la liberté ou non des passagers de se déplacer librement sur le navire. Cependant, nous ne pouvons pas tirer de conclusions ou d'inférences solides à partir de ces seules données.

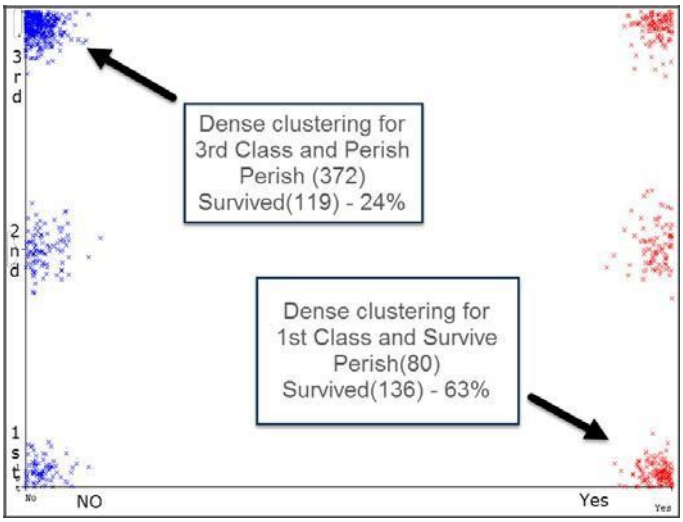


Fig. 3. Visualisation simple par la méthode K-means – Survivants vs classe

G. Survivants vs groupe d'âge

Notre normalisation des données a arbitrairement réparti les données des passagers en différents groupes nominaux. Parmi ces groupes, sans surprise, les adultes âgés de 20 à 49 ans figuraient parmi ceux qui ont péri. La figure 4 ci-dessous ne présente pas un regroupement visuel aussi important que les deux précédentes (sexe ou âge). Notre approche des tranches d'âge était une généralisation. Une analyse plus approfondie pourrait être effectuée. avec des groupes d'âge plus granulaires ou définis généalogiquement afin de mettre en évidence d'autres relations potentielles qui pourraient exister.

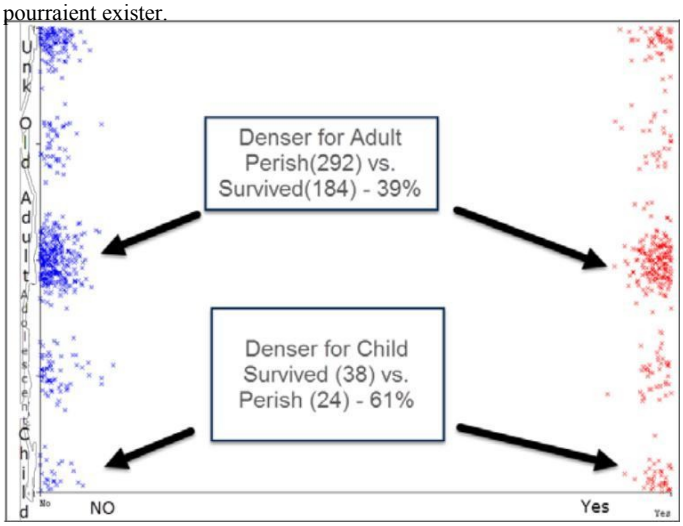
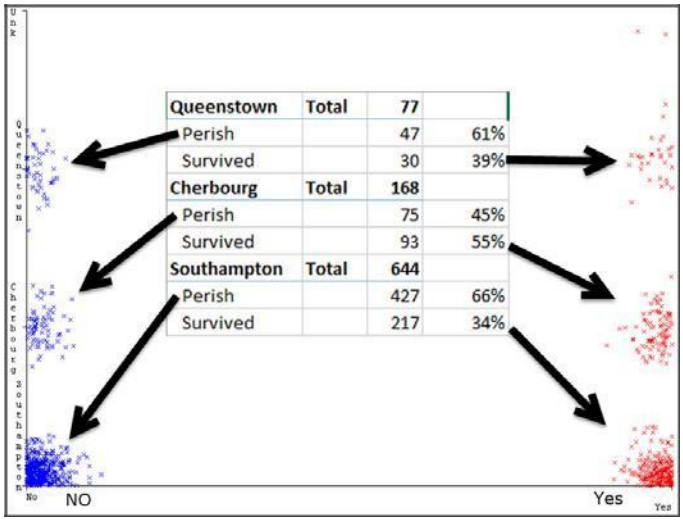


Fig. 4. Visualisation simple K Means – Survivants vs. Groupe d'âge

H. Survivants vs embarqués

Enfin, l'analyse a permis d'identifier que le point d'embarquement des passagers était également un indicateur du taux de survie, bien que moins significatif. Ce qui n'a pas été fait, c'est l'association du point d'embarquement avec la classe de cabine – par exemple, les passagers de 3e classe ont-ils principalement embarqué à Southampton ? Ceci est illustré dans la figure 5 ci-dessous.



Queenstown	**Total**	**77**	
Perish		47	61%
Survived		30	39%
Cherbourg	**Total**	**168**	
Perish		75	45%
Survived		93	55%
Southampton	**Total**	**644**	
Perish		427	66%
Survived		217	34%

Fig. 5. Visualisation simple K Means – Survivants vs embarqués

Enfin, l'indication sur le diagramme de regroupement de la classe des passagers montre une forte association entre la classe et le point de départ. Ainsi, ces deux caractéristiques semblent liées d'une certaine manière et ne sont probablement pas indépendantes.

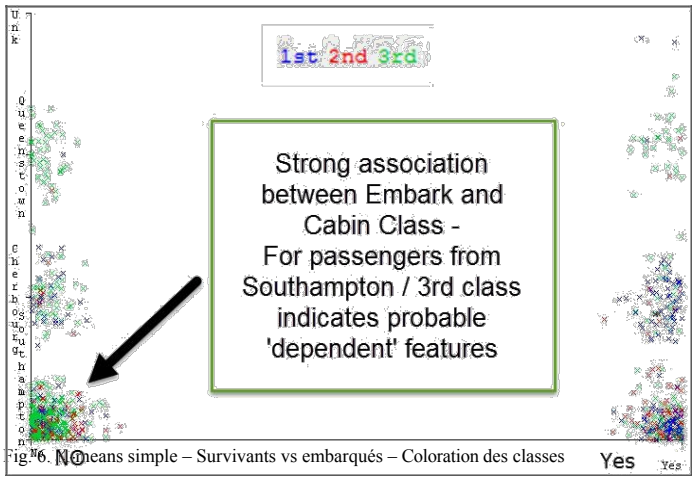


Fig. 6. K-means simple – Survivants vs embarqués – Coloration des classes

IV. Observations finales

Le sexe était clairement le facteur le plus significatif dans le jeu de données pour le taux de survie. De plus, le classificateur J48, utilisant le jeu de données de test, a permis de classer correctement environ 81 % des cas. Par rapport au concours Kaggle au moment de la rédaction de cet article, cela plaçait le modèle à la 43e place environ.

V. Résumé

Bien que puissant, l'outil Weka nécessite de convertir les données dans un format plus convivial afin de faciliter son utilisation et le fonctionnement des approches de classification. Cela a été une bonne expérience d'apprentissage pour l'équipe. Au départ, nous avions choisi les statistiques du baseball, mais nous avons rapidement été submergés par le nombre d'attributs et la taille des ensembles de données. La conversion des données de type numérique en classificateurs étant une tâche fastidieuse, nous avons abandonné les statistiques du baseball et cherché un autre ensemble de données. Weka et les algorithmes nécessitaient des valeurs nominales pour les classificateurs plutôt que des valeurs numériques.

Nous avons découvert l'ensemble de données de Kaggle et, grâce à une manipulation simple, nous avons pu obtenir un ensemble de données tout à fait compatible au format ARFF (format natif de Weka) qui fonctionnait bien et fournissait des résultats assez significatifs démontrant quelles classes de passagers avaient le plus grand impact sur la survie.

VI. Recherches futures

L'ensemble de données utilisé représentait un sous-ensemble ou un ensemble de données « test » utilisé pour le concours Kaggle. Avec l'ensemble de données complet, le modèle peut être validé et certaines des mêmes conclusions ou relations vérifiées. En outre, il convient d'examiner certaines des autres dépendances de classification croisée, telles que la classe de cabine et le lieu d'embarquement, afin d'éliminer les classificateurs inutiles.

VII. Références

[1] GE, « Flight Quest Challenge », Kaggle.com. [En ligne]. Disponible sur : https://www.gequest.com/c/flight2-main. [Consulté le 13 décembre 2013].

[2] « Titanic : Machine Learning from Disaster », Kaggle.com. [En ligne]. Disponible sur : https://www.kaggle.com/c/titanic-gettingStarted. [Consulté le 13 décembre 2013].

[3] Wiki, « Titanic ». [En ligne]. Disponible sur : http://en.wikipedia.org/wiki/Titanic. [Consulté le 13 décembre 2013].

[4] Kaggle, Data Science Community, [En ligne]. Disponible sur : http://www.kaggle.com/ [Consulté le 13 décembre 2013]

[5] Weka 3 : logiciel d'exploration de données en Java, [en ligne]. Disponible à l'adresse : http://www.cs.waikato.ac.nz/ml/weka/ [Consulté le 13 décembre 2013]

[6] Algorithme C4.5, Wikipédia, Wikimedia Foundation, [En ligne]. Disponible sur : http://en.wikipedia.org/wiki/C4.5_algorithm, [Consulté le 13 décembre 2013]

VIII. Annexes

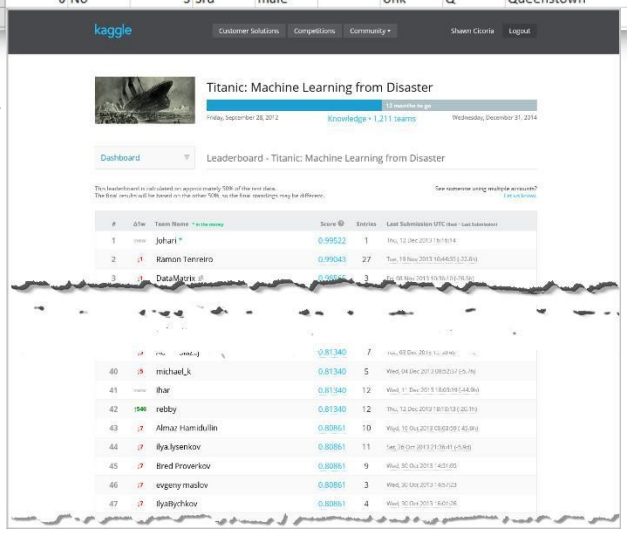
A. *Exemple de données provenant de Kaggle – Ensemble de données initial*

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, Mr. C	male	22	1	0	A/5 2117	7.25		S
3	2	1	1	Cumings, Mrs	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, M	female	26	0	0	STON/O2	7.925		S

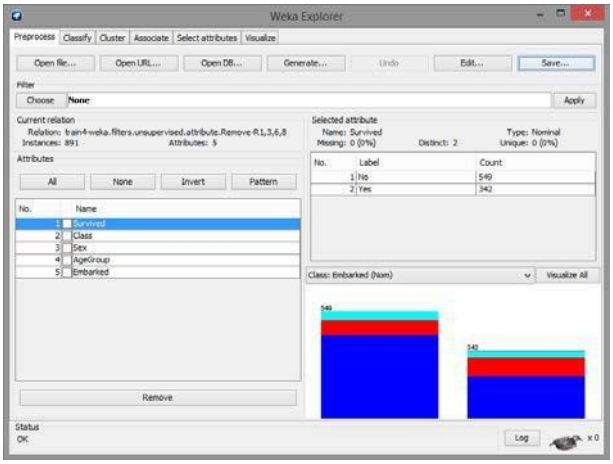
B. *Ensemble de données normalisé basé sur les données Kaggle*

	A	B	C	D	E	F	G	H	I	J
1	Passenger	Survived	Pclass	Class	Sex	Age	AgeGroup	Ecode	Embarked	
2	1	No	3	3rd	male	22	Adult	S	Southampton	
3	2	Yes	1	1st	female	38	Adult	C	Cherbourg	
4	3	Yes	3	3rd	female	26	Adult	S	Southampton	
5	4	Yes	1	1st	female	35	Adult	S	Southampton	
6	5	No	3	3rd	male	35	Adult	S	Southampton	
7	6	No	3	3rd	male		Unk	Q	Queenstown	
8										

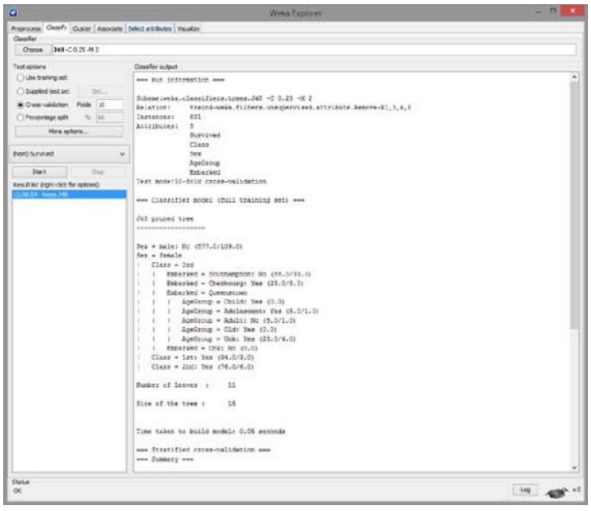
C.

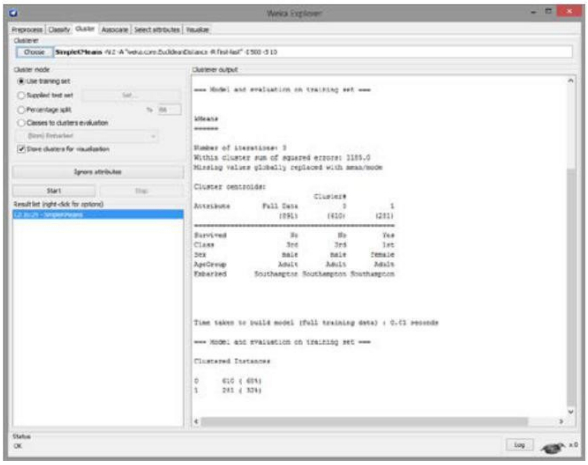


D. Écrans Weka – Vue principale du prétraitement



E. Écrans Weka – Vue de classification J48





F. Ensemble de données normalisé au format ARFF

```
@relation 'train4-weka.filters.unsupervised.attribute.Remove-R1,3,6,8'

@attribut Survived {Non,Oui}
@attribut Class {1er,2e,3e}
@attribut Sex {masculin,féminin}
@attribut AgeGroup {Enfant,Adolescent,Adulte,Vieux,Inconnu} @attribut
Embarked {Southampton,Cherbourg,Queenstown,Inconnu}

@data Non,3e,homme,adulte,Southampton
Oui,1re,femme,adulte,Cherbourg
```

G. Résultat du classificateur J48 (3 parties)

```
=== Informations d'exécution === Schéma :
weka.classifiers.trees.J48 -C 0,25 -M
2 Relation :
   train4-
weka.filters.unsupervised.attribute.Remove-R1,3,6,8
Instances : 891 Attributs :
5 Survivants
                    Classe
                    Sexe
                    Groupe
                    d'âge
                    Embarqué

Mode de test : validation croisée 10 fois
```

```
=== Modèle de classification (ensemble
d'apprentissage complet) élagué J48
Sexe = femme
|    Classe = 3e
|    |    Embarquée = Southampton : Non (88,0/33,0)
|    |    Embarqué = Cherbourg : Oui (23,0/8,0)
|    |    Embarqué = Queenstown
|    |    |    Groupe d'âge = Enfant : Oui (0,0)
|    |    |    Groupe d'âge = Adolescent : Oui
(5,0/1,0)
|    |    |    Groupe d'âge = Adulte : Non (5,0/1,0)
|    |    |    Groupe d'âge = Personnes âgées : Oui
                (0,0)
|    |    |    Groupe d'âge = Inconnu : Oui (23,0/4,0)
|    |    Embarqué = Inconnu : Non (0,0)
|    Classe = 1ère : Oui (94,0/3,0)
|    Classe = 2e : Oui (76,0/6,0)


Nombre de feuilles :      11
Taille de l'arbre :       15
Temps nécessaire à la construction du modèle
: 0,05 seconde

=== Validation croisée stratifiée ===
```

=== Validation croisée stratifiée ===
=== Résumé ===

Instances correctement classées	722	
81,0325 %		
Instances incorrectement classées	169	
18,9675 %		
Statistique kappa	0,571	
Erreur absolue moyenne	0,2911	
Erreur quadratique moyenne	0,385	
Erreur absolue relative	61,5359	%
Erreur quadratique relative	79,1696	%
Nombre total d'instances	891	

=== Précision détaillée par classe ===

	Taux TP	Taux de FP	Précision	Rappel
	0,953	0,418	0,785	0,953
	0,582	0,047	0,884	0,582
Moyenne pondérée	0,81	0,276	0,823	0,81

Mesure F	Aire ROC	Classe
0,861	0,783	Non
0,702	0,783	Oui
0,8	0,783	

=== Matrice de confusion ===

a	b	<-- classé comme
523	26 |	a = Non
143	199 |	b = Oui

H. Résultats du regroupement simple par la méthode K-means

=== Informations d'exécution ===
Schéma : weka.clusterers.SimpleKMeans
-N 2 -A « weka.core.EuclideanDistance - R first-last » –I 500 -S 10
Relation : train4-weka.filters .unsupervised.attribute .Remove-R1,3,6,8 Instances : 891
Attributs : 5 Survivants
de classe Sexe
Groupe d'âge
Embarqué

Mode de test : évaluation sur les données d'apprentissage

=== Modèle et évaluation sur l'ensemble d'apprentissage === kMeans
======
Nombre d'itérations : 3
Somme des erreurs quadratiques au sein du cluster : 1185,0 Valeurs manquantes remplacées globalement par la moyenne/le mode Centres de gravité des clusters :

Attribut	Données complètes	Groupe n° 0	1
	(891)	(610)	(281)
Survivants	Non	Non	Oui
Classe	3e	3e	1
Sexe	masculin	Masculin	Femme
Tranche d'âge	Adulte	Adulte	Adulte
Embarqué	Southampton	Southampton	Southampton

Temps nécessaire à la construction du modèle (données d'apprentissage complètes) : 0,01 seconde
=== Modèle et évaluation sur l'ensemble d'apprentissage === Instances regroupées

0	610 (68 %)
1	281 (32 %)