**Proiect Introducere in Knowledge Management**

1. Tehnologia folosita: WEKA

<http://www.cs.waikato.ac.nz/ml/weka/>

Set de date: Supermarket

<https://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/vote.arff>

Nume: Alban Marius - Ionut

Grupa: 10LF471

Facultatea: Matematica si Informatica

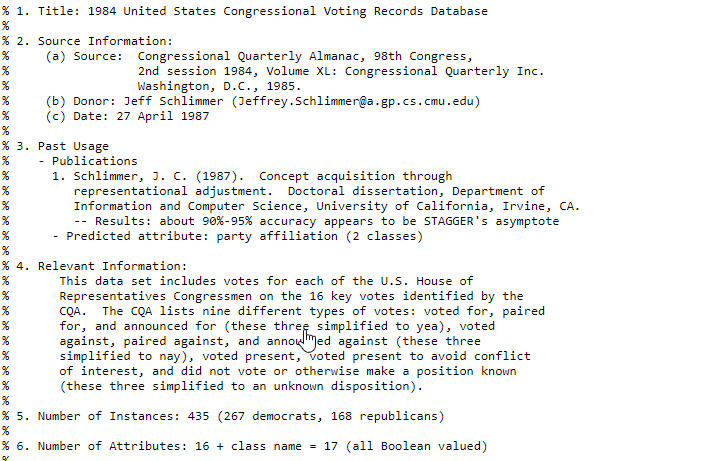
Specializarea: IAG

An : II

Acest set de date include voturi pentru fiecare dintre Camerele SUA din Congres reprezentate pe cele 16 voturi-cheie identificate de CQA. CQA enumeră nouă tipuri diferite de voturi: votat pentru, asociaza pentru și anunțate pentru (cele trei simplificate pentru yea), au votat împotriva, s-au legat împotriva și au anunțat împotriva (aceste trei simplificate la nay

Numarul instantelor este de 435 (267 democrati , 168 republicani).

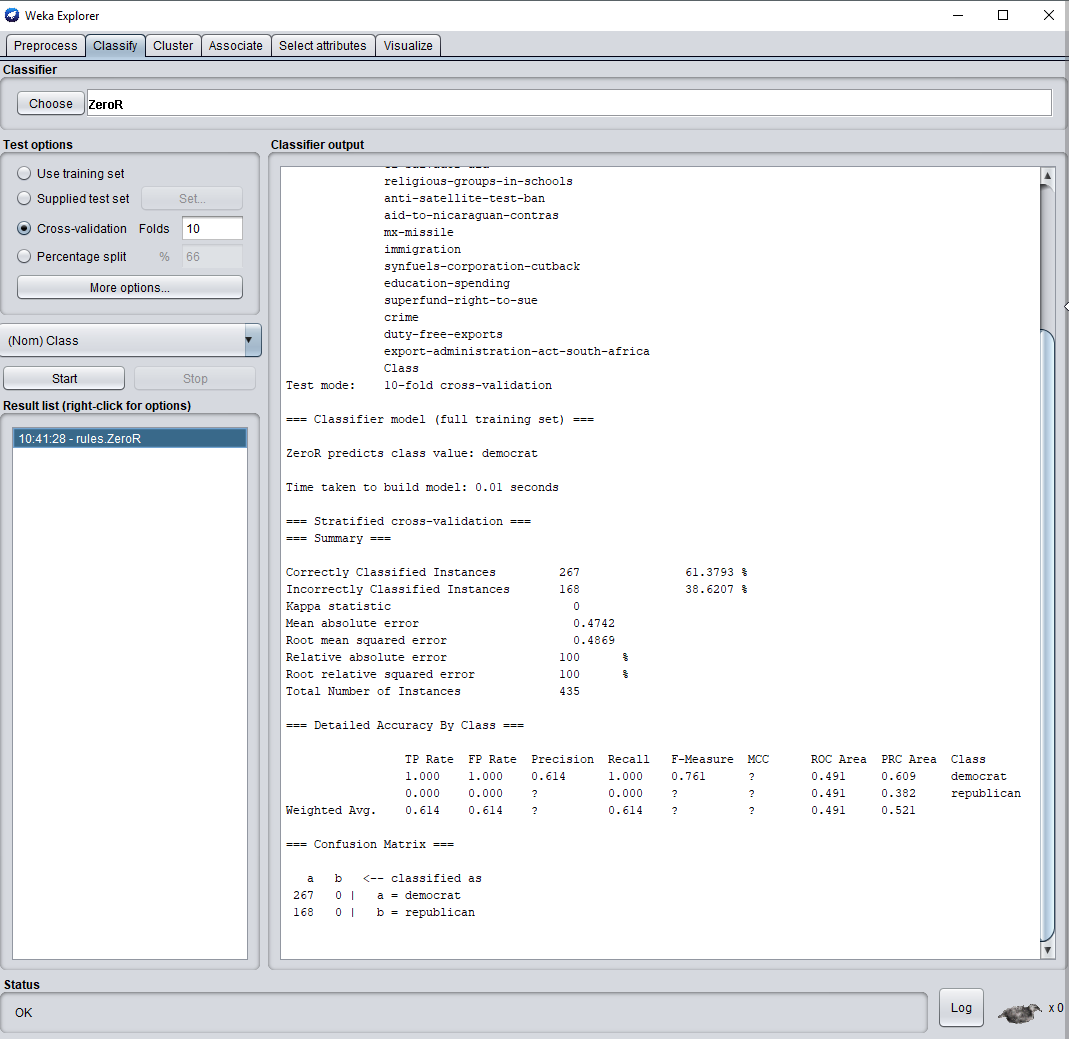
Numarul atributelor de 16,daca este inclusa si clasa 17. (valori boolean)



Setul de date poate fi folosit in clasificari si clusterizare,dar si pentru asociere si vizualizare.

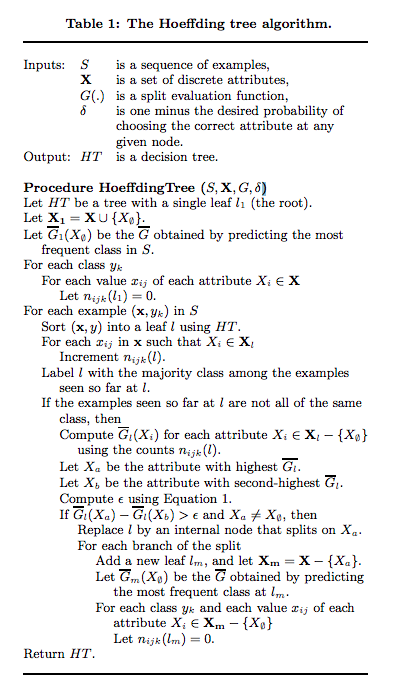
Weka este o colectie de algoritmi de invatare pentru data mining. Algoritmii pot fi aplicati fie direct pe un set de date. Weka contine instrumente pentru preprocesarea datelor, clasificare, regresie, reguli de asociere si pentru vizualizare. De asemenea, este potrivit pentru dezvoltarea de noi scheme de invatare. Este utilizat in cercetare, educatie si in cadrul aplicatiilor.

1. **Clasificare.**

Clasificarile în WEKA sunt modelele pentru estimarea cantităților nominale sau numerice. Schemele de învățare disponibile în WEKA includ arbori și liste de decizie, clasificatori bazați pe instanțe,mașini vectoriale de suport, perceptronuri cu mai multe straturi si regresie logistică. ****

**Altoritmul „Hoeffding tree”**

Algoritmul „Hoeffding tree” este un algoritm de inductie incrementata, fiind un arbore de decizie, care este capabil să invete din fluxurile masive de date, presupunând că exemplele generatoare de distributie nu se schimba in timp. „Hoeffding tree” exploateaza faptul ca un mic esantion poate fi adesea suficient pentru a alege un atribut optim de divizare.

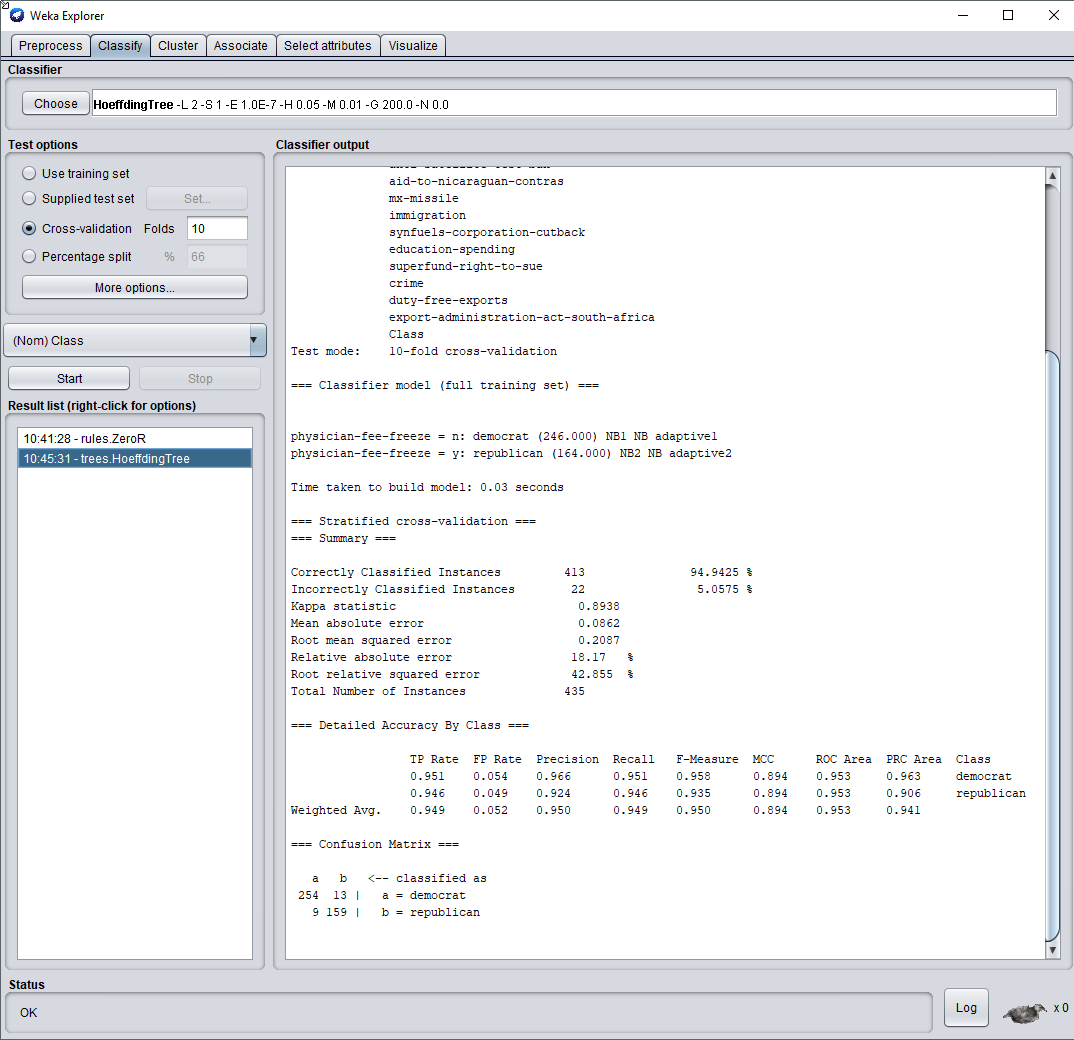


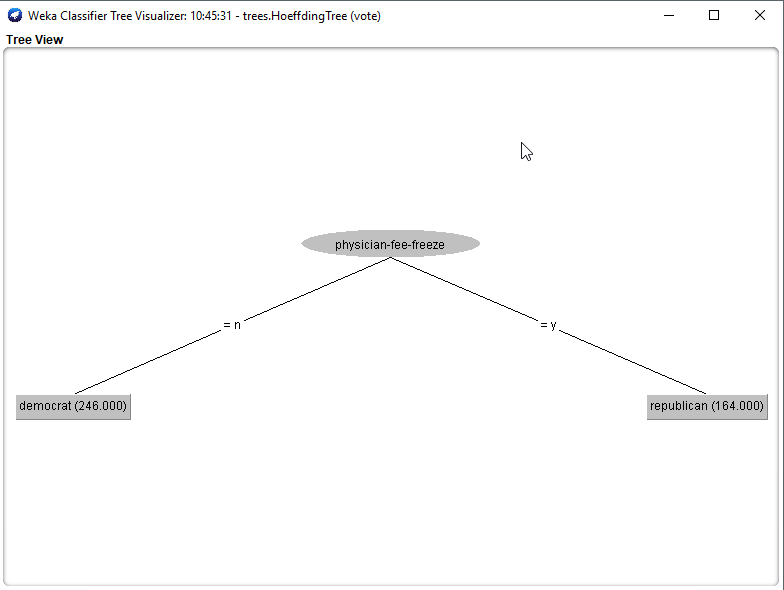
Am folosit algoritmul pe setul de date folosind optiunea use training set.

Numarul de instante clasificate corect este de 413.

Procentul de acuratete este de 94.9425 %.Timpul pentru a testa aceasta clasificare a fost de 0,03 s.

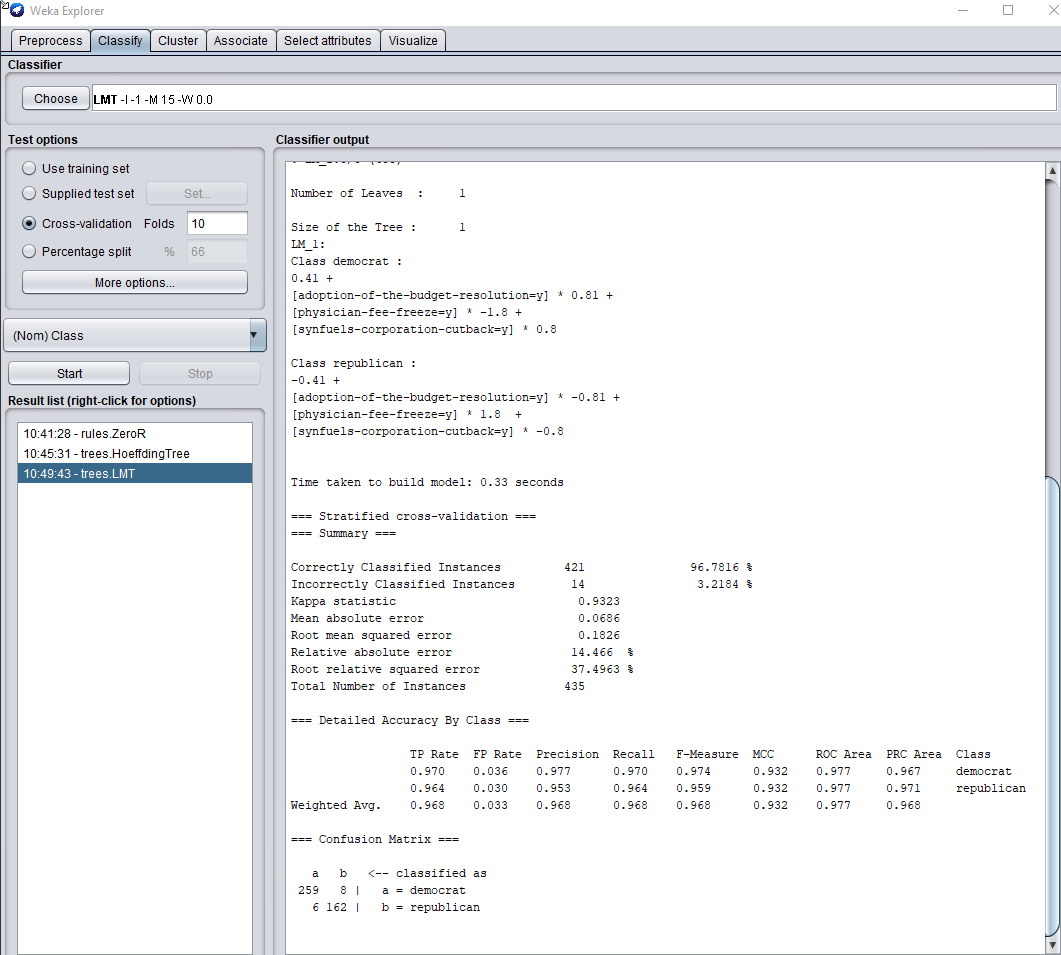
Arborele construit in urma algoritmului „Hoeffding tree” este:





1. **Algoritmul LMT(Arborele modelului logistic)**

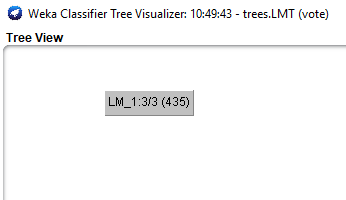
Arborele modelului logistic(LMT) se bazează pe ideea anterioară a unui arbore model: un arbore de decizie care are modele de regresie liniară la frunzele sale pentru a furniza un model de regresie liniară pe piesă (unde arborii decizionali obișnuiți cu constante la frunzele lor ar produce un model constant pe bucată). Algoritmul de bază de inducție LMT utilizează validarea încrucișată pentru a găsi un număr de iterații LogitBoost care nu depășesc datele folosite.



Numarul de instante clasificate corect este de 421.

Procentul de acuratete este de 96.7816 %.

Arborele construit in urma algoritmului „Hoeffding tree” este:



**Comparatie:**

Dupa folosirea celor doi algoritmi,se observa cateva diferente.

Una dintre acestea este ca algoritmul LMT are instante clasificate corect 421, iar algoritmul „Hoeffding tree” 413.

Procentul de acuratete este de 94 % la algoritmul „Hoeffding tree” si la cel de-al doilea algoritm de 96 % LTM .

Concluzia este ca Algoritmul LTM este mult mai efficient decat LTM.

1. **Clusterizare.**

.

Un cluster reprezinta o colectie de obiecte care sunt „asemanatoare”intre ele si sunt ”diferite” fata de obiectele apartinand acestor clustere.

Clusterele se clasifica in doua categorii:

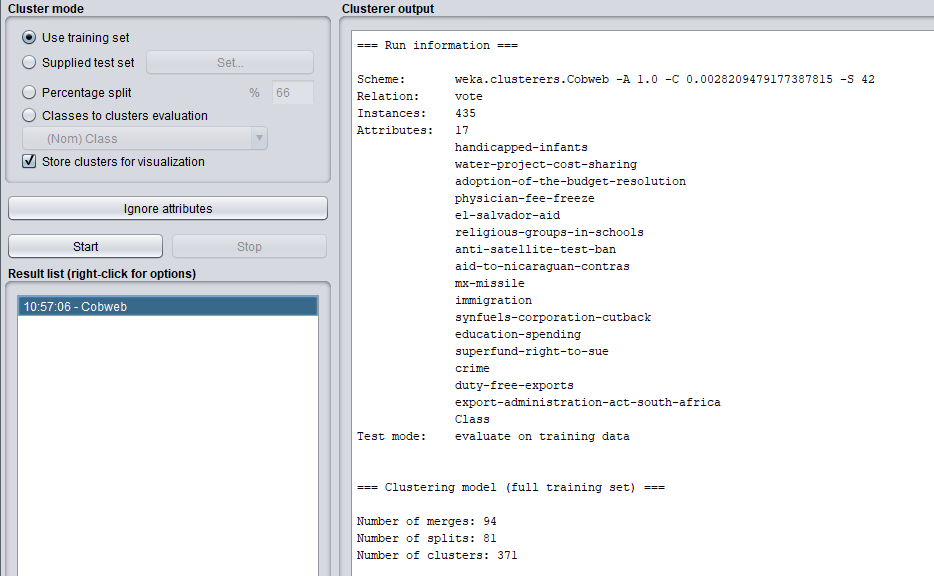
* Lineari(k-means,fuzzy c-means)
* Non-lineari(algoritmul bazat pe MST,kernel k-means)

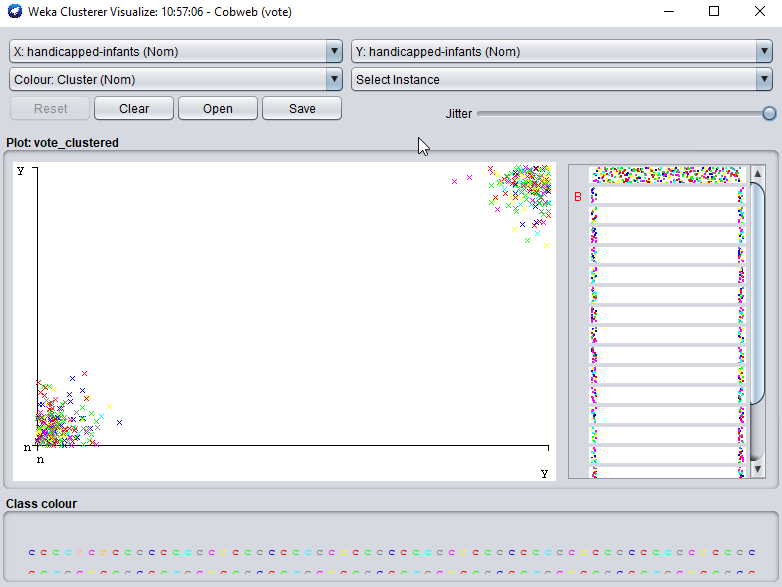
1. **Algoritmul Cobweb**

COBWEB este un sistem incrementat pentru gruparea conceptuală ierarhică. Algoritmul organizează treptat observațiile într-un arbore de clasificare. Fiecare nod dintr-un arbore de clasificare reprezintă o clasă (concept) și este etichetat printr-un concept probabilistic care rezumă distribuțiile atribut-valoare ale obiectelor clasificate sub nod. Acest arbore de clasificare poate fi folosit pentru a prezice atributele lipsă sau clasa unui obiect nou.

Există patru operațiuni de bază care utilizează COBWEB pentru construirea arborelui de clasificare.

* Îmbinarea a două noduri
* Împărțirea unui nod(  Un nod este împărțit prin înlocuirea lui cu copiii acestuia)
* Introducerea unui nou nod( Se creează un nod corespunzător obiectului inserat în arbore)
* Trecerea unui obiect în jos în ierarhie(Apelarea efectivă a algoritmului COBWEB se face pe obiect și pe subarborele înrădăcinat în nod)

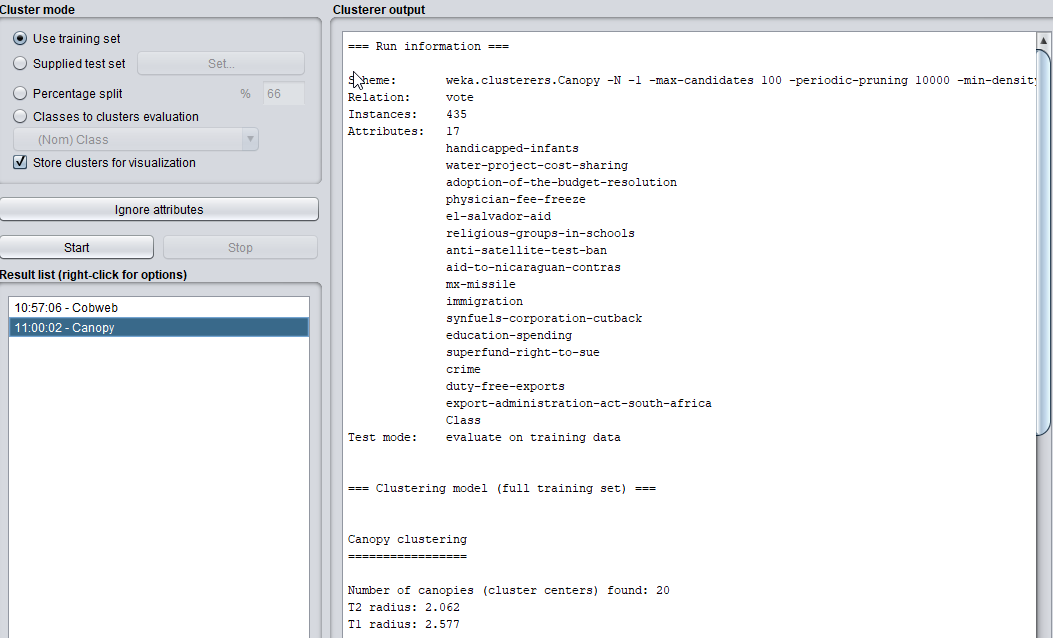




1. **Algoritmul Canopy**

Algoritmul de Canopy este un algoritm de pre-clusterizare nesupravegheat.

Este adesea folosit ca pas de preprocesare a algoritmului K-means sau a algoritmului de grupare ierarhica .



**Comparatie** intre cei doi algoritmi de clusterizare:

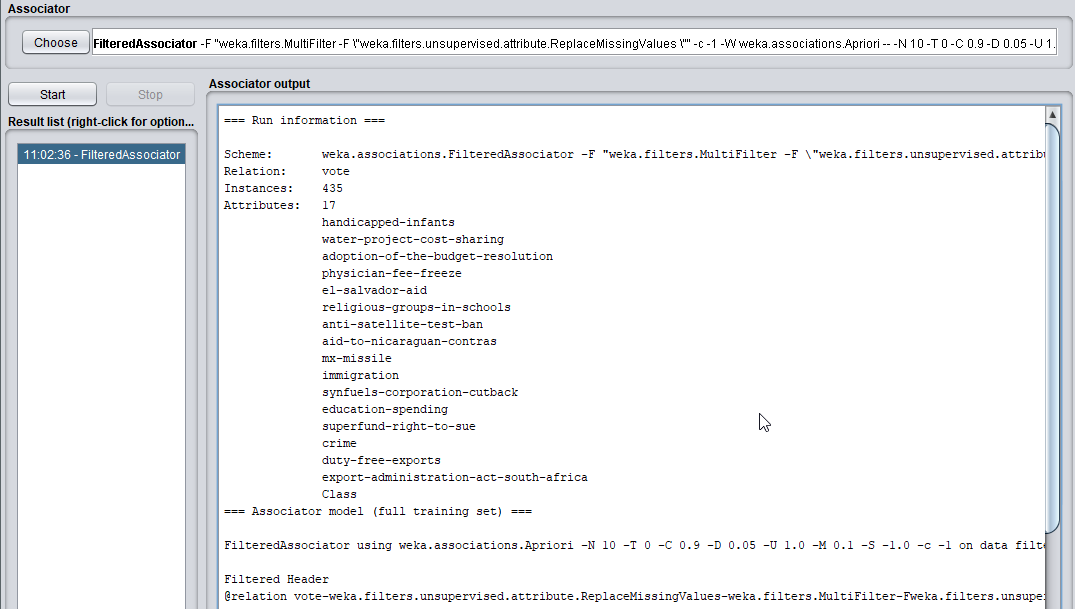
Algoritmul Cobweb a clusterizat 371 de instante,iar algoritmul Canopy 20.

1. **Asociere.**

Asocierea este o functie de extragere a datelor care descopera probabilitatea co-apari-tiei elementelor intr-o colectie.Relatiile dintre elemente care apar simultan sunt exprimate ca reguli de asociere.

FilteredAssociator

Clasa pentru rularea unui asociator arbitrar pe date care au fost trecute printr-un filtru. Ca si asociatul, structura filtrului se bazeaza exclusiv pe datele de antrenament si instantele de testare vor fi procesate de filtru fara a schimba structura lor.

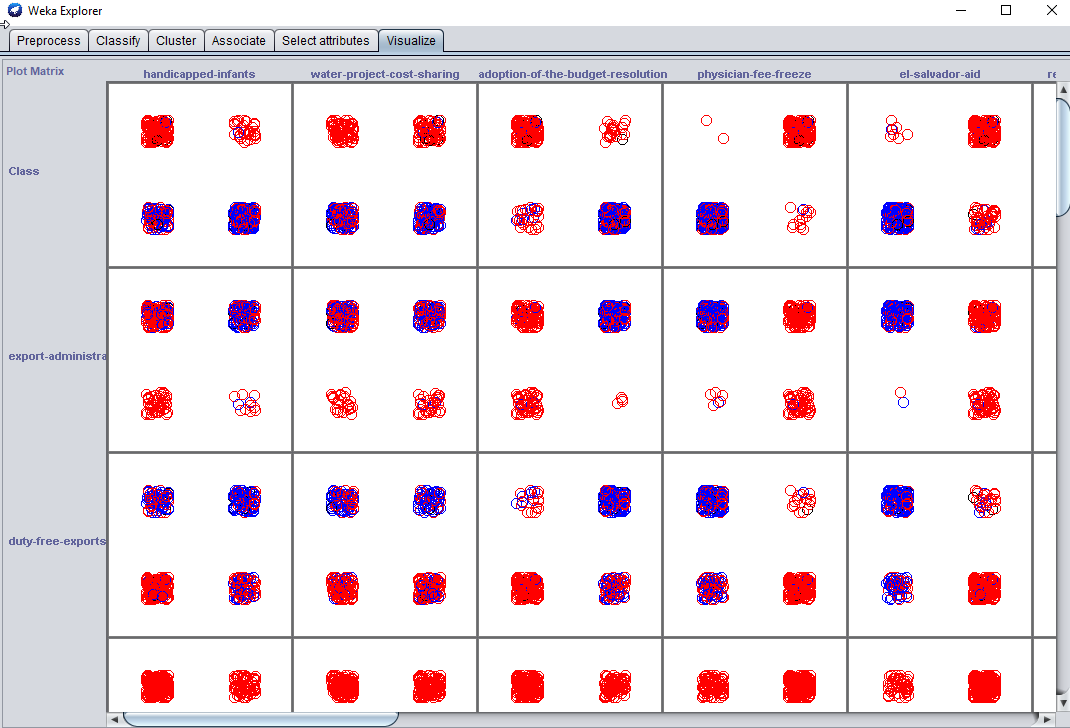


1. **Vizualizare.**

Vizualizarea tuturor atributelor



Vizualizare dupa voturi pentru crime.



**Bibliografie**

<https://www.researchgate.net/figure/Hoeffding-Tree-Algorithm_fig2_224255453>

<https://en.wikipedia.org/wiki/Logistic_model_tree>

<https://en.wikipedia.org/wiki/Cobweb_(clustering)>

<https://en.wikipedia.org/wiki/Canopy_clustering_algorithm>

<https://wiki.pentaho.com/display/DATAMINING/FilteredAssociator>