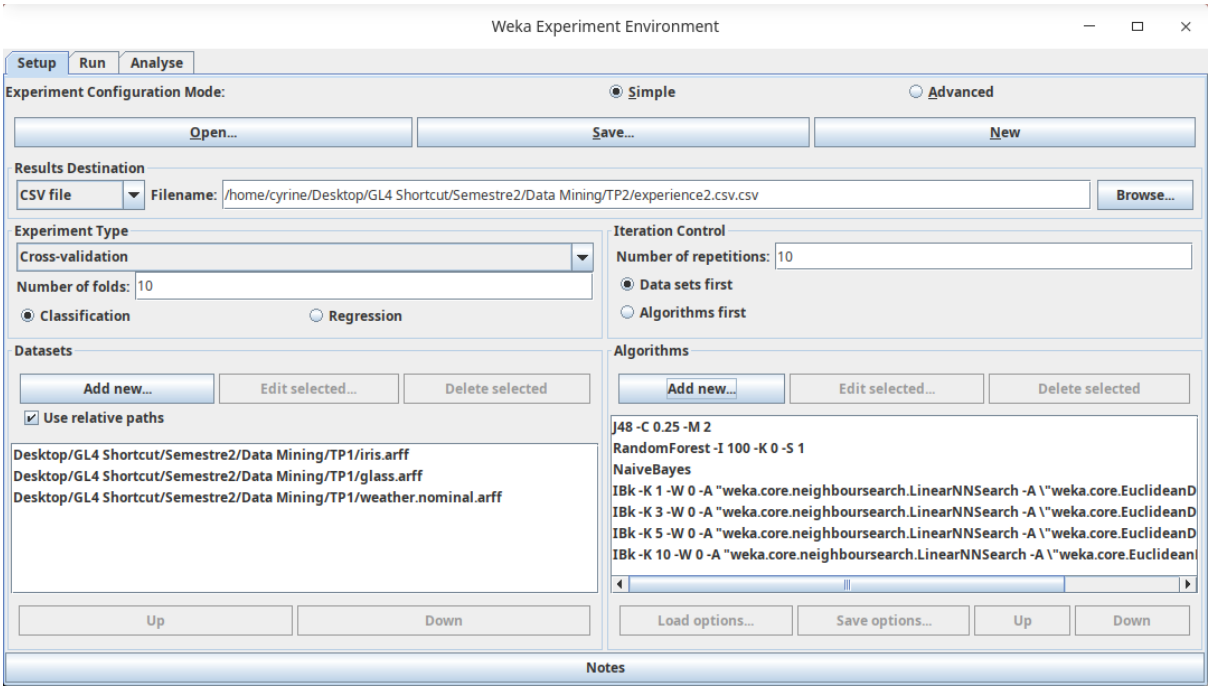
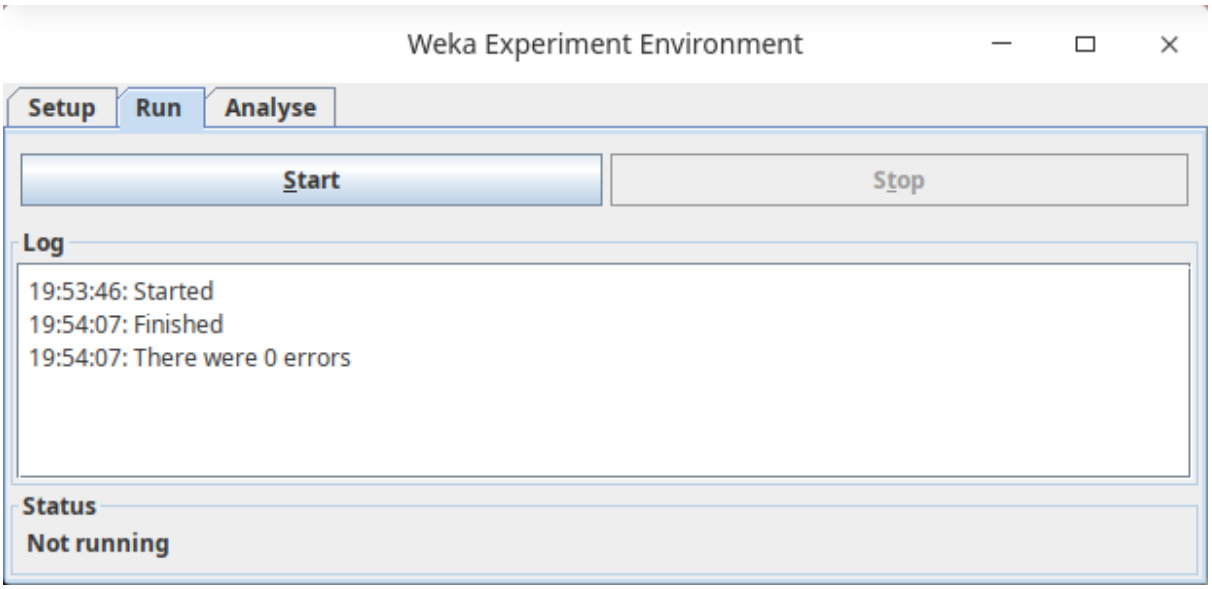
Essia Ben Hamida - Cyrine Zaouali - Souha Ben Hassine - GL4/G2

Compte Rendu TP2 Data Mining

Les colonnes utilisées pour comparer les classifieurs sont : Accuracy, Recall, précision & f-value.
Les classifieurs : J48, RandomForest, Naive Bayes, IBK k=1, IBK k=3, IBK k=5, IBK k=10
Les Data Sets : IRIS, Glass et Weather

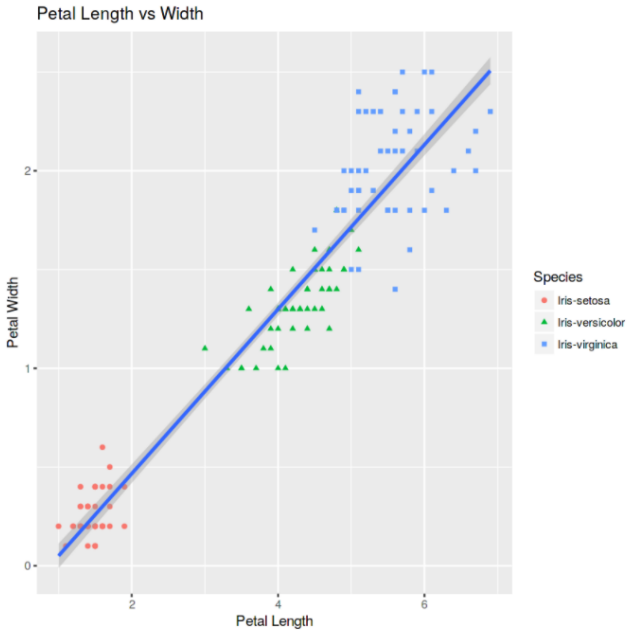
Setup :





Dataset 1: IRIS

Classifieur	Correct %	Incorrect %	Recall	Précision	F-measure
J-48	94,73333333	5,266666667	1	0,98	0,9888888889
RandomForest	94,73333333	5,266666667	1	1	1
Naive Bayes	95,53333333	4,466666667	1	1	1
IBK k=1	95,4	4,6	1	1	1
IBK k=3	95,2	4,8	1	1	1
IBK k=5	95,73333333	4,266666667	1	1	1
IBK k=10	95,73333333	4,266666667	1	1	1



Conclusion : Pour la dataset IRIS on remarque que les classifieurs IBK k=5 et IBK k=10 nous donnent les meilleurs résultats par rapport aux pourcentages de la correction (accuracy).

On remarque que les trois classes sont distribuées => Notre dataset est clean. Ceci explique les résultats élevés obtenus pour tous les classifieurs. Le IBK a donné de très bons résultats car tous les voisins ou presque sont de la même classe.

Dataset 2: Glass

Classifieur	Correct %	Incorrect %	Recall	Précision	F-measure
J-48	67.6255	32.3744	0.7142857	0.70460714	0.6957006
RandomForest	79.9264	20.07359	0.86857	0.793498	0.829338
Naive Bayes	49.44588	50.55411	0.744285	0.478470	0.576208
IBK k=1	69.95022	30.04978	0.752857	0.711473	0.724778
IBK k=3	70.01515	29.98485	0.832857	0.6526890	0.724192
IBK k=5	66.04329	33.95671	0.832857	0.633345	0.713299
IBK k=10	63.25541	36.74459	0.862857	0.607944	0.825466

Conclusion : Pour la dataset Glass on remarque que le classifieur Random Forest nous donne les meilleurs résultats par rapport aux pourcentages de la correction (accuracy).
On remarque que même pour les autres métriques (recall, precision et f-mesure) Random Forest reste le meilleur. Random Forest étant un classifieur ensembliste, il utilise plusieurs modèles en même temps. Ceci explique le résultat trouvé. Ce dataset contient 213 instances ce qui explique

Dataset 3: Weather

Classifieur	Correct %	Incorrect %	Recall	Précision	F-measure
J-48	47.5	52.5	0.42	0.355	0.376
RandomForest	69	31	0.74	0.63	0.66
Naive Bayes	57.5	42.5	0.64	0.53	0.566
IBK k=1	61.5	38.5	0.63	0.535	0.566
IBK k=3	70.5	29.5	0.82	0.675	0.72
IBK k=5	71	29	0.9	0.71	0.773
IBK k=10	70	30	0.9	0.7	0.766

Conclusion : Pour la dataset Weather on remarque que le classifieur IBK k=5 nous donne les meilleurs résultats par rapport aux pourcentages de la correction (accuracy). On remarque que même pour les autres métriques (recall, precision et f-mesure) Weather reste le meilleur. Le dataset Weather contient

10 instances donc K=5 est juste au milieu ce qui donne la moyenne. Pour K= 10, le nombre d'instances, le résultat n'est plus précis de même pour K=1.

Conclusion générale :

Selon les résultats obtenus, la meilleure performance en termes de taux de précision de la classification globale est donnée par les algorithmes Random Forest de la catégorie ensembliste et IBk de la catégorie des Lazy classificateurs. L'algorithme IBK est un classificateur k-Nearest-Neighbor, qui s'est avéré donner de bons résultats en termes de classification.
Pour conclure, le choix du classifieur dépend du dataset, est-il concentré sur un résultat? Est-il nettoyé? nombre de classes..