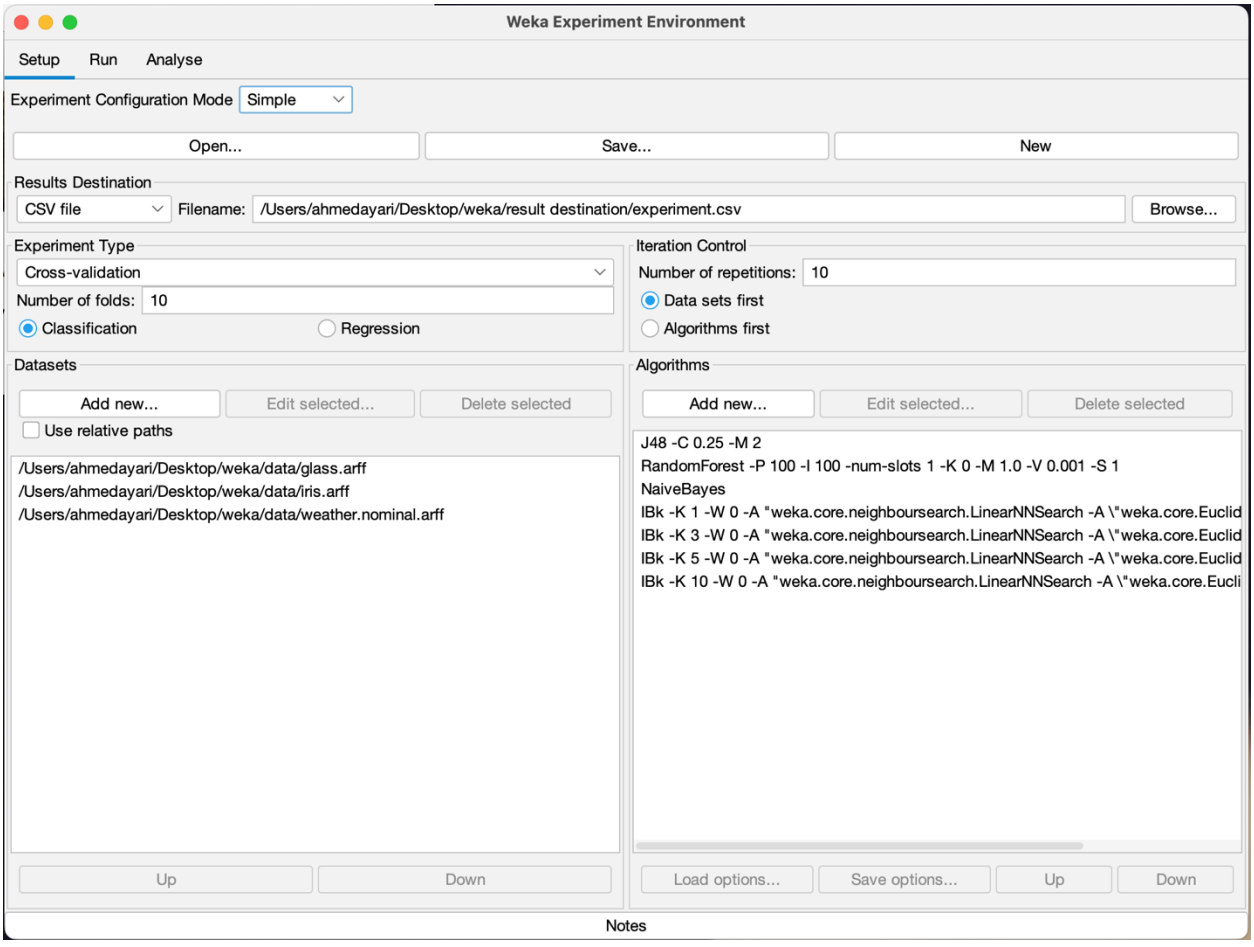
TP2 – Data Mining: Introduction to Weka Experimenter

Realized by:
Ahmed Ayari,
Oussema Zouaghi,
Moahmed Karim Mallouli

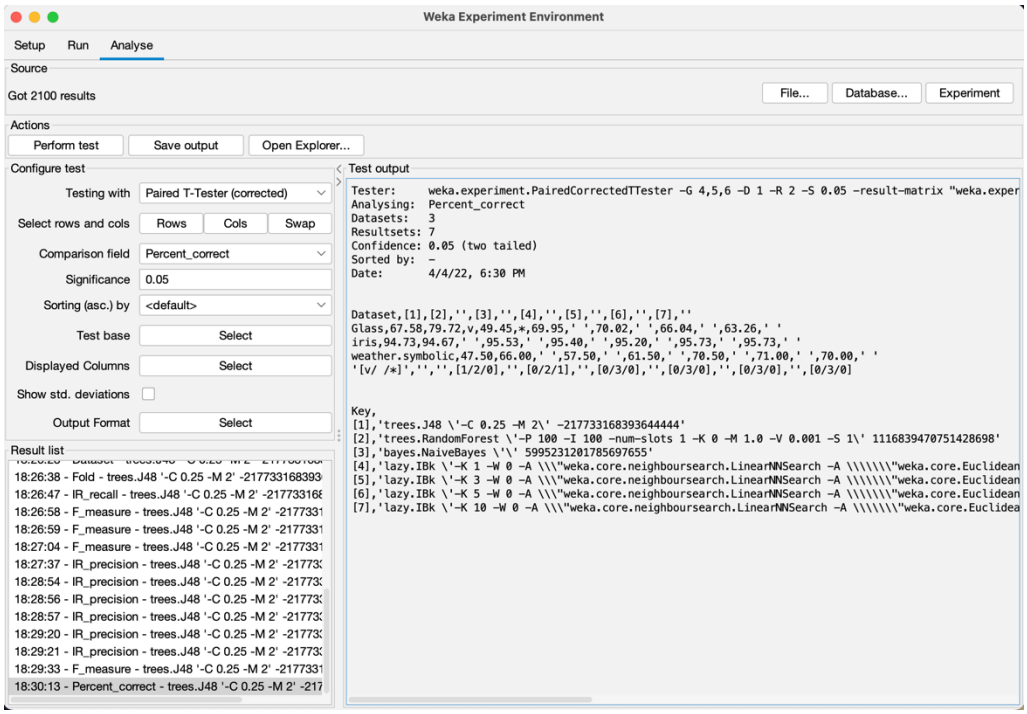
Experience description:

- Models: J48, RandomForest, Naive Bayes, IBK k=1, IBK k=3, IBK k=5, IBK k=10
- Datasets:
  - Iris: A dataset containing observation of a lot of iris flowers and their races
  - Glass: dataset containing observations about glass types and their chemical composition
  - Weather: dataset containing observations of the weather status each corresponding to the fact that tennis (or other sport) went outside to practice or not
- Comparison attributes : Accuracy, Recall, Precision and f-value

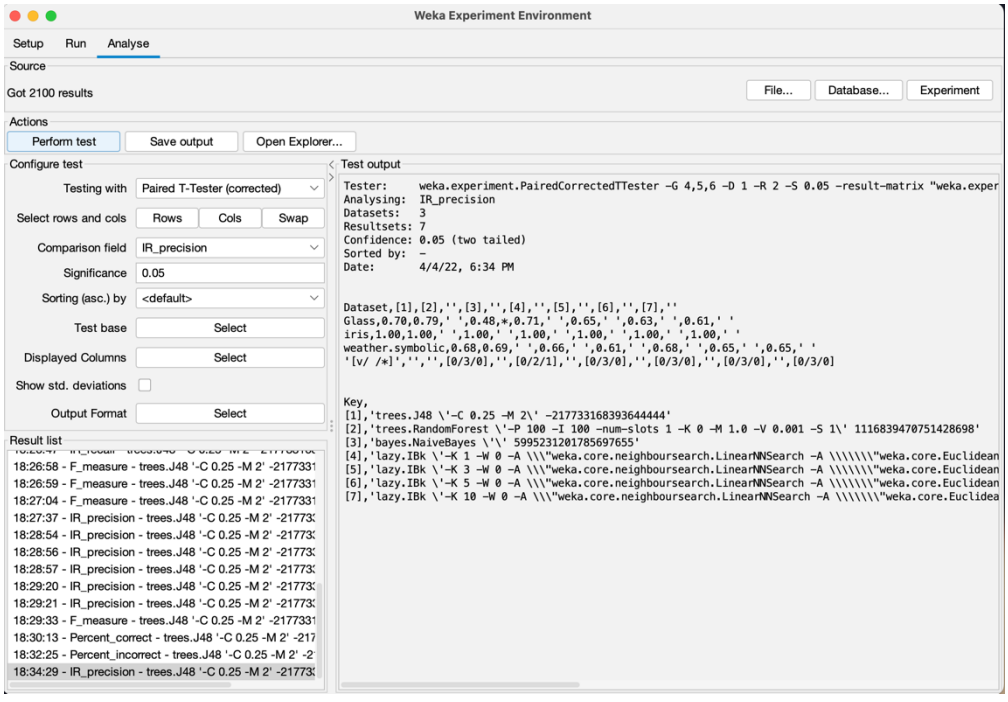


Test Results:

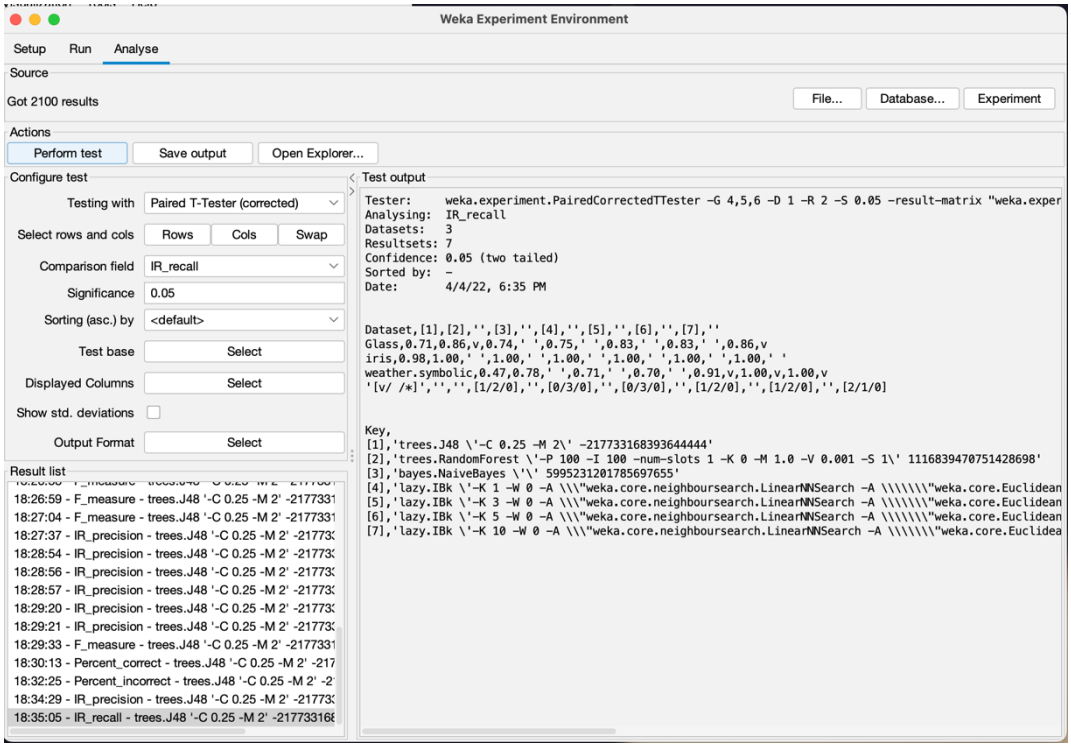
- Percent correct (Accuracy):



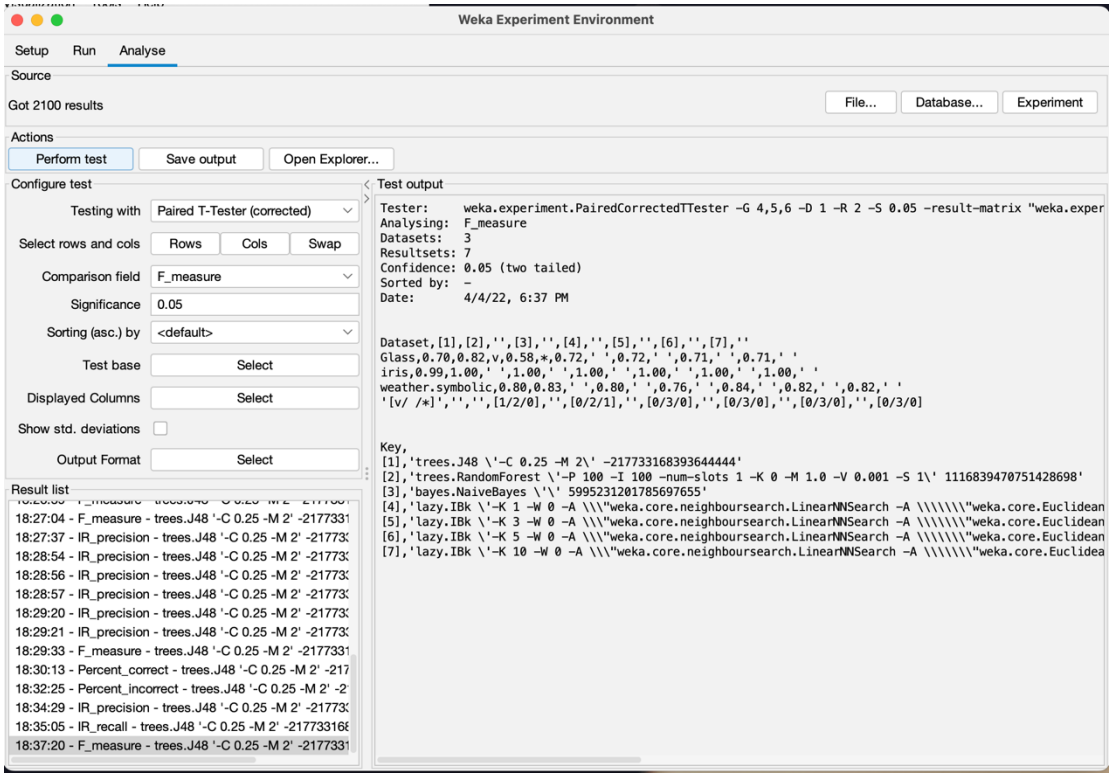
- Precision:



- Recall



- F-value (calculated from the precision and recall of the test)



Results Analysis:

- For accuracy attribute:
  - Glass dataset: best result with Random Forest Model
  - Iris dataset: best result with IBK model with K=5 and K=10
  - Weather dataset: best result with IBK model with K=5

- For precision attribute:
  - Glass dataset: best result with J48 Model
  - Iris dataset: all results=1
  - Weather dataset: best result with Random Forest model

- For recall attribute:
  - Glass dataset: best result with Random Forest and IBK with k=10 Models
  - Iris dataset: all results=1
  - Weather dataset: best result with IBK model with K=5 and K=10

- For f-value attribute:
  - Glass dataset: best result with Random Forest Model
  - Iris dataset: all results=1
  - Weather dataset: best result with IBK model with K=3

Conclusions:

- Glass dataset: the model should predict the type of glass created depending on the matter used to create it.
  The best model here is the one with higher precision rate (less false positive prediction), having higher precision will result in lesser costs.
  ⇨ Best model: Random Forest
- Iris dataset: based on the context, there are no losses caused by false negative or false positive (not a metric to choose the best model).
  Accuracy would be the right attribute to choose the corresponding model, and based on the results above:
  ⇨ Best Model: IBK with K=5 and K=10
- Weather dataset: model is used to determine whether a player will go to practice or not based on the weather conditions.
  If this model is used in training spaces (stadium preparation), false negative would have more impact than false positive, a model with less false negative is more suitable: lesser false negative implies higher recall.
  ⇨ Best models: IBK with K=5 and K=10