Data Mining with Weka

Association-rule mining, supervised classification & unsupervised clustering across retail and scientific datasets

Author: Erich Assuncao | **Context:** Graduate coursework — COMP 682, Data Mining | **Toolkit:** Weka 3 | **2025**

OVERVIEW

Summary

This project applies three of the core data-mining families — **association rule mining**, **classification**, and **clustering** — inside the Weka workbench, working across one retail transaction dataset and two scientific benchmark datasets. It moves from *unsupervised pattern discovery* (which products are bought together), through *supervised prediction* (classifying iris species from measurements), to *structure discovery* (finding the natural groupings inside iris and glass data).

Each module pairs a disciplined experimental setup — dataset preprocessing, parameter tuning, and quantitative evaluation — with a plain-language interpretation of *what the patterns mean* and *how they could be used*. The strongest results from the original graduate submission are curated here and reframed as an applied case study in choosing, tuning, and reading data-mining models.

METHODS & TOOLS

Approach

A single consistent workflow was applied to every module: **select** a dataset, **preprocess** it (format handling, class-label removal, normalization), **tune** algorithm parameters, **evaluate** with the metric appropriate to the task, and **interpret** the output.

- **Association & market-basket** — Apriori, evaluated by support, confidence, and lift.
- **Classification** — NaiveBayesSimple, MultilayerPerceptron, and J48 (C4.5), evaluated by accuracy and Cohen's kappa under both training-set and 5-fold cross-validation.
- **Clustering** — SimpleKMeans, EM, and HierarchicalClusterer, evaluated by within-cluster sum of squares (WCSS), log-likelihood, and dendrogram structure.

Dataset	Instances	Attributes	Role in the study
supermarket	~4,600 baskets	Boolean item flags	Market-basket / association rules
Iris	150	4 numeric + 1 class	Classification & clustering
Glass	214	9 numeric + 1 class	Clustering (chemical composition)

MODULE 1

Association Rule Mining — Market-Basket Analysis

Apriori was run over the supermarket transaction dataset to surface products that co-occur in the same basket. A preprocessing detail mattered: a nominal-to-binary conversion prevented Apriori from running, so the raw data was reloaded in its native format, which the algorithm accepted directly. Two parameter regimes were then explored.

Run A — ranked by confidence

With minimum support 0.05 and minimum confidence 0.6, all ten top rules resolved to the same consequent — **bread and cake** — at confidence ≈ 0.96–0.97 and lift ≈ 1.34. Strong but generic: a signal that the basket-size attribute (`total`) was dominating.

```
baking needs=t, biscuits=t, pet foods=t, cheese=t, milk-cream=t, fruit=t, total=high  (242)
   ==> bread and cake=t  (234)     conf:(0.97)   lift:(1.34)   lev:(0.01)   conv:(7.54)
```

Run B — ranked by lift, basket-size removed

Switching the metric to **lift** and dropping the `total` attribute exposed much stronger, more specific associations — lift ≈ **2.5**, i.e. co-occurrences roughly 2.5× more likely than chance — linking bakery items with frozen and produce goods.

```
biscuits=t, frozen foods=t, tissues-paper prd=t, vegetables=t  (795)
   ==> bread and cake=t, fruit=t, total=high  (475)     conf:(0.60)   lift:(2.50)
```

Run	Ranking metric	Top confidence	Top lift	Character of rules
A	Confidence	0.97	1.34	Generic — all → "bread & cake"
B	Lift (no `total`)	0.60	2.54	Specific — bakery ↔ frozen/produce

What the store could do with it

- **Highest-confidence rule (0.60):** shoppers already holding biscuits, frozen foods, tissues, and vegetables have a 60% chance of adding bread & cake and fruit. → Advertise a "*fruit & bakery snack pack*" at the produce-section entrance.
- **Highest-lift rule (2.54):** baskets with bread & cake, biscuits, sauces, and vegetables are 2.5× more likely to also pick up frozen foods and fruit. → An *end-cap display* near the deli/produce aisle bundling a frozen sampler with fruit.
- **General uses:** cross-promotions, shelf placement, and targeted coupon issuance to lift cross-sales.

MODULE 2

Classification — Predicting Iris Species

The Iris dataset (150 flowers described by four petal/sepal measurements across three species) was used to compare three classifiers with different inductive biases. Each was scored on the training set and under 5-fold cross-validation, measuring both accuracy and Cohen's kappa.

Classifier	Training accuracy	5-fold CV accuracy	Kappa (CV)
NaiveBayesSimple	96.00%	96.00%	0.94
MultilayerPerceptron	98.67%	96.00%	0.94
J48 (C4.5)	98.00%	96.00%	0.94

Reading the results. All three converge to 96% under cross-validation — Iris is an "easy," nearly linearly separable dataset whose features are close to conditionally independent given the class, which is exactly why even a simple Naïve Bayes model matches the neural network. The perceptron and decision tree fit the training data more tightly (98–98.7%) but gain *no* cross-validation advantage, which quantifies mild overfitting rather than better generalization.

Deployment recommendation. For real-world use I would favour NaiveBayesSimple or the pruned J48 tree for their speed, interpretability, and robustness, reserving the MultilayerPerceptron for genuinely nonlinear problems where enough data exists to keep overfitting in check.

MODULE 3

Clustering — Discovering Natural Structure

Three clustering algorithms were applied to two datasets after removing class labels and normalizing all features to [0,1] with a shared Euclidean distance. The goal was to see how partitional, model-based, and hierarchical methods each describe the same data.

Iris

Algorithm	Clusters	Cluster sizes	Quality metric
SimpleKMeans (k=3)	3	61 / 50 / 39	WCSS = 6.9981
EM (auto)	5	28 / 35 / 42 / 22 / 23	Log-likelihood = 2.3233
Hierarchical (avg, k=3)	3	50 / 67 / 33	dendrogram-based

Glass

Algorithm	Clusters	Cluster sizes	Quality metric
SimpleKMeans (k=3)	3	161 / 51 / 2	WCSS = 31.3461
EM (auto)	2	139 / 75	Log-likelihood = 10.5237
Hierarchical (complete, k=3)	3	184 / 3 / 27	dendrogram-based

Reading the results. On Iris, k-means and hierarchical both recover three groupings that align with the true species, and the dendrogram confirms the well-known pattern that *Setosa* separates first while the other two species merge later. EM, allowed to choose its own component count, splits into five — surfacing finer sub-structure within a species. On Glass, EM finds a strong *binary* division in composition (two clusters), while k-means and hierarchical expose small outlier clusters (down to a two- and three-instance singleton) that flag unusual glass samples.

Takeaway. No single algorithm wins universally: k-means gives crisp, fast partitions; EM performs automatic model selection and adapts to cluster shape; hierarchical reveals nested, multi-level relationships that flat methods cannot. The right choice depends on whether the task needs clean partitions, automatic structure inference, or an exploratory hierarchy.

REFLECTION

Limitations & Next Steps

- **Attribute-removal trade-off.** Dropping `total` removed the basket-size signal — it killed generic "high-value" rules but may also have hidden patterns unique to large transactions. A discretized spend metric could be reintroduced selectively.
- **Rule homogeneity.** Even after tuning, top rules still orbit the bakery ↔ frozen/produce axis. Higher support thresholds or alternative consequents (dairy, snacks) would diversify the actionable clusters.
- **Modest lift & parameter sensitivity.** Lift under 3 is meaningful but moderate, and results depend on the support/confidence choice; a grid search or conviction / chi-square ranking would tighten which associations are both significant and commercially useful.
- **Clean benchmark data.** Iris and Glass are small, tidy datasets — the near-perfect classification and clean clusters would not transfer directly to noisier, higher-dimensional real-world data.

TAKEAWAYS

Skills Demonstrated

Association rule mining · market-basket analysis · Apriori (support / confidence / lift) · supervised classification · Naïve Bayes, neural networks & decision trees · cross-validation & Cohen's kappa · overfitting diagnosis · unsupervised clustering · k-means, EM & hierarchical methods · data preprocessing & normalization · the Weka workbench · translating model output into business-relevant insight.

> *Adapted from a graduate data-mining assignment (COMP 682) completed as part of the Master of Science in Computing and Information Systems at Athabasca University. The datasets, analyses, and quantitative results shown here are from the original academic submission; this document curates and reframes them as an applied portfolio case study.*