

School of Computing and Information Systems
The University of Melbourne
COMP30027 Machine Learning (Semester 1, 2021)
Workshop: Week 4

| ID | A (°C) | B (mm) | C (hPa) | CLASS |
|----|--------|--------|---------|-------|
| 1 | 22.5 | 4.6 | 1021.2 | AUT |
| 2 | 16.7 | 21.6 | 1027.0 | AUT |
| 3 | 29.6 | 0.0 | 1012.5 | SUM |
| 4 | 33.0 | 0.0 | 1010.4 | SUM |
| 5 | 13.2 | 16.4 | 1019.5 | SPR |
| 6 | 14.9 | 8.6 | 1016.4 | SPR |
| 7 | 18.3 | 7.8 | 995.4 | WIN |
| 8 | 16.0 | 5.6 | 1012.8 | WIN |

1. What is **Discretisation**, and where might it be used?
 - (i). Summarise some approaches to **supervised** discretisation.
 - (ii). Discretise the above dataset according to the (unsupervised) methods of **equal width**, **equal frequency**, and **k-means** (breaking ties where necessary).
2. Find the (sample) **mean** and (sample) **standard deviation** for the attributes in the above dataset:
 - (i). In its entirety, and;
 - (ii). For each individual class;
 - (iii). How could we use this information when building a classifier over this data?
3. How is **holdout** evaluation different to **cross-validation** evaluation? What are some reasons we would prefer one strategy over the other?
4. A **confusion matrix** is a summary of the performance of a (supervised) classifier over a set of development (“test”) data, by counting the various instances:

| | | Actual | | | |
|------------|----------|----------|----------|----------|----------|
| | | <i>a</i> | <i>b</i> | <i>c</i> | <i>d</i> |
| Classified | <i>a</i> | 10 | 2 | 3 | 1 |
| | <i>b</i> | 2 | 5 | 3 | 1 |
| | <i>c</i> | 1 | 3 | 7 | 1 |
| | <i>d</i> | 3 | 0 | 3 | 5 |

- (i). Calculate the classification **accuracy** of the system. Find the **error rate** for the system.
- (ii). Calculate the **precision**, **recall**, **F-score** (where $\beta = 1$), **sensitivity**, and **specificity** for class *d*. (Why can't we do this for the whole system? How can we consider the whole system?)