School of Computing and Information Systems
The University of Melbourne
COMP90049
Knowledge Technologies (Semester 1, 2018)
Workshop exercises: Week 11

1. For the following dataset:

ID	*Outl*	*Temp*	*Humi*	*Wind*	PLAY
TRAINING INSTANCES					
A	s	h	h	F	N
B	s	h	h	T	N
C	o	h	h	F	Y
D	r	m	h	F	Y
E	r	c	n	F	Y
F	r	c	n	T	N
TEST INSTANCES					
G	o	c	n	T	?
H	s	m	h	F	?

   Classify the test instances using a Decision Tree:

   (a) Using the Information Gain as a splitting criterion

   (b) Using the Gini Index as a splitting criterion

2. What is **bagging**, in the context of **Decision Trees**?

   (a) What is a **Random Forest**?

   (b) What advantages does a Random Forest have, with comparison to a (deterministic) Decision Tree model, or a bag of Decision Trees?

3. For the following dataset:

apple	*ibm*	*lemon*	*sun*	CLASS
TRAINING INSTANCES				
4	0	1	1	FRUIT
5	0	5	2	FRUIT
2	5	0	0	COMPUTER
1	2	1	7	COMPUTER
TEST INSTANCES				
2	0	3	1	?
1	0	1	0	?

   (a) Using the **Euclidean distance** measure, classify the test instances using the 1-NN method.

   (b) It is also possible to use a similarity measure for k-NN, rather than a distance measure: using the **Cosine similarity**, classify the test instances using the 3-NN method.