TDS2101 Introduction to Data Science

Tutorial 4

1. Splitting your data into different sets is an essential procedure for building predictive models.

   (a) What is the purpose of having a test set and a training set when building models?

   (b) In some cases, data can also be split into 3 sets: training, validation and test. What is the rationale of having the additional validation set?

2. What is the advantage of using **cross-validation** as opposed to custom splitting of dataset into train and test sets?

3. Using application examples, differentiate **supervised learning**, **unsupervised learning** and **reinforcement learning**.

4. What classification performance metrics can be used in the case of imbalanced data?

   Provide an example/scenario of why this is necessary.

5. The following confusion matrix of a model has been obtained from evaluated test data of patients diagnosed with diabetes from a particular community.

		Predicted class	
		Y	N
Actual class	Y	99	42
	N	8	?

   (a) Given that the model obtained an accuracy of 80%, how many true negative samples were there?

   (b) One of the two types of falsely predicted results seems to be much higher than the other. What kind of real-world implications can we expect if this model is used?

   (c) Is test set is *class balanced?* Why do you say so?