Introduction to Data Mining
Individual Project I: Classification

Due: 10/12/2015 - 10:30am

Project Description

In this project, you are asked to apply different classification methods on the given datasets. You should submit:

- A *Readme.txt* file, explaining how to run your script. (If there is any ambiguity in working with your functions)
- A report (no longer than 3 pages) showing:
  - Details of each classification method for the datasets. (e.g. confusion matrices, plots)
  - A final table for each dataset, having these columns: <*Method*, *Accuracy*, *Precision*, *Recall*, and *F-Measure*>
  - Your conclusion.
- A .R script, containing every function definintion. By running this script, we should be able to get your reported results.

Submit your files as a .zip file: **FirstnameLastname_UFID.zip**.

Datasets

First, you need to divide your dataset into training set (80%) and test set (20%). Randomly select samples for training set and test set, but for result reproduciblity, use a deterministic seed:

```
set.seed(1890) # Use your first and last two digits of UFID
```

Iris Flower Dataset

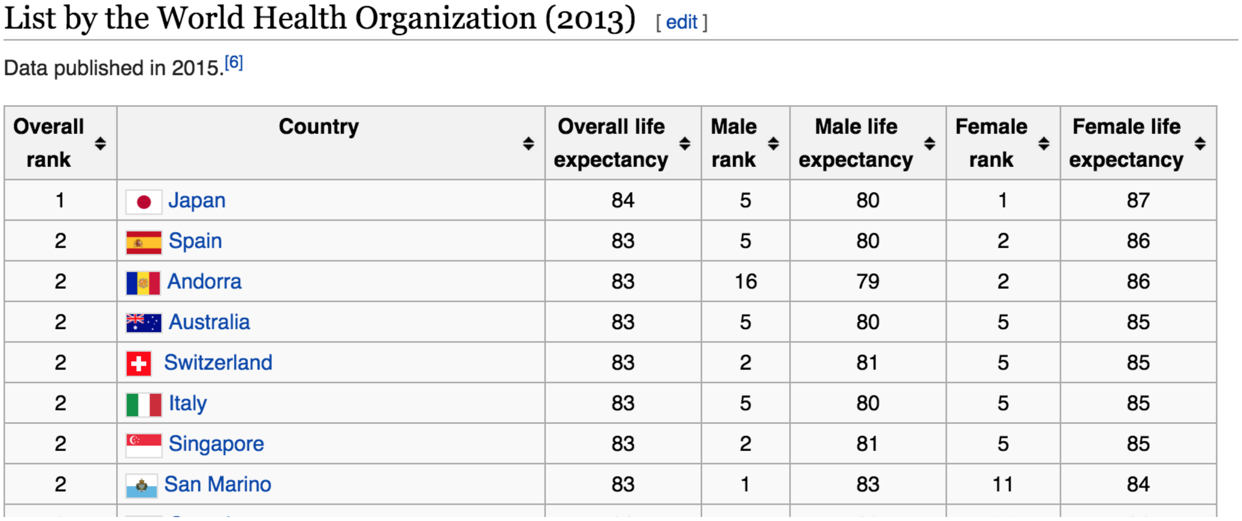
This dataset contains 50 samples from each of three species of Iris flower and each sample has four features: length and width of the sepals and petals. This dataset is available in R *datasets* package.

```
install.packages("datasets", repos="http://cran.rstudio.com/")
# Or just simply use
# install.packages("datasets")

library(datasets)
data("iris")
```

Life Expectancy Dataset

This dataset can be obtained from wikipedia.org/wiki/List_of_countries_by_life_expectancy. See figure below:



List by the World Health Organization (2013) [edit]

Data published in 2015.[6]

Overall rank	Country	Overall life expectancy	Male rank	Male life expectancy	Female rank	Female life expectancy
1	Japan	84	5	80	1	87
2	Spain	83	5	80	2	86
2	Andorra	83	16	79	2	86
2	Australia	83	5	80	5	85
2	Switzerland	83	2	81	5	85
2	Italy	83	5	80	5	85
2	Singapore	83	2	81	5	85
2	San Marino	83	1	83	11	84

To use this dataset, add a *Continent* column. It is not provided in the given link, but you can easily find continent of every country by a simple search. This additional column plays as the class label role. Preserve other columns as well.

Classification Methods

For any mentioned classfication method, try to find the right package and method. If you cannot find it, you have to write your own method. You may want to tune the classification method by providing the right arguments, based on your understanding on the dataset in use. You should clearly mention why you choose such parameters (for instance, the number of nearest neighbors in *kNN* method).
Use these classifications for the above-mentioned datasets. Do not forget: the reported results (final tables) should be about your classification performances on the **test set**. Performance on the training should be brought in as the details of each method.

- Decision Tree
  - RIPPER
  - C4.5
  - Oblique
- Naive Bayes
- k-Nearest Neighbor (kNN)

Script

Along with your report, you should submit your script (.R file) too. Your script should contain every line of code to be executed to get the final results.

```
# Your function definition

# Example
```

```
myC45 <- function(...) {
     # learns a fit based on the given training set
     # returns a fit
}
myC45Predict <- function(...) {
     # for each sample in test set, it predicts the label.
     # returns the required values (accuracy, precision, ...)
}
# Other function definitions

# Reading datasets and dividing into training and test sets
divideDataset <- function(...){
     set.seed(1890)
     # returns:
     # dataset$training
     # and dataset$test
}

# Calling the functions and getting the results
```

If we are not able to run your script (by following your *Readme.txt* file), you will not receive any points. Your results should be consistent with what you report. If there is a specific variable that holds the final values, you should mention it in your *Readme.txt* so we know how to check your results.