# Preface

For this assignment you will need to use Weka. If you can’t use the lab computer to finish it on time, then you can download it from the following link for your operating system. Download the developer version: <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>

You can also come to the lab to finish the assignment. The easiest would be to use the lab computers.

# Assignment

There are two datasets available for this assignment as separate files: labor data and segment-challenge data. The description of these datasets is available within each of them. You can view the contents of the data files by right clicking on them and opening them in Wordpad.

To answer the following questions create a **separate Word document**:

1. Apply K-means clustering algorithm on the labor data and segment-challenge data by changing the number of cluster from 2 to 10 by incrementing 1. Also select the *Cluster mode* on the *Cluster* tab in Weka to *Classes to clusters evaluation*. Now, every time you change the number of clusters for each dataset, record the percentage of ***incorrectly*** clustered instances from the output (at the bottom). Find out the best number of clusters for each data. The best number of clusters would be the one which has the minimum percentage of **incorrectly** clustered instances. Report the best number of clusters and the percentage of incorrectly clustered instances for each dataset in a separate Word document
   * To learn how to use K-means in Weka, take a look at the instructions of the lab.

**Incorrectly clustered instances : 13.0 22.807 %**

Incorrectly clustered instances : 21.0 36.8421 %

Incorrectly clustered instances : 29.0 50.8772 %

Incorrectly clustered instances : 31.0 54.386 %

Incorrectly clustered instances : 33.0 57.8947 %

Incorrectly clustered instances : 32.0 56.1404 %

Incorrectly clustered instances : 38.0 66.6667 %

Incorrectly clustered instances : 34.0 59.6491 %

Incorrectly clustered instances : 36.0 63.1579 %

Dataset 2:

Incorrectly clustered instances : 1073.0 71.5333 %

Incorrectly clustered instances : 837.0 55.8 %

Incorrectly clustered instances : 636.0 42.4 %

Incorrectly clustered instances : 588.0 39.2 %

Incorrectly clustered instances : 573.0 38.2 %

**Incorrectly clustered instances : 501.0 33.4 %**

Incorrectly clustered instances : 597.0 39.8 %

Incorrectly clustered instances : 620.0 41.3333 %

Incorrectly clustered instances : 718.0 47.8667 %

1. Can you explain the reason for getting the best number of clusters that you got in task 1? Write only a short answer of one to two sentences in the Word document.

**For both labour and segment challenge datasets, we got best accuracies of 77.193 % and 66.6% at k=2 and k=7 respectively as compared to other values of k. And hence, the reason for getting getting best number of clusters is high accuracy rate.**

1. Submit your Word document containing the results of tasks 1-2 to Blackboard.