Instructions:

- This assignment must be completed individually.
- Plagiarism will not be tolerated. Reference where appropriate.
- Answer all of the questions.
- Generate a pdf document with your answers and submit it to the upload link provided on the course website.

Plagiarism Policy

The Department of Computer Science considers plagiarism as a serious offence. Disciplinary action will be taken against students who commit plagiarism. Plagiarism includes copying someone else’s work without consent, copying a friend’s work (even with consent) and copying textual material from the Internet. Copying will not be tolerated in this course. For a formal definition of plagiarism, the student is referred to http://www.ais.up.ac.za/plagiarism/index.htm (from the main page of the University of Pretoria site, follow the Library quick link, and then click the Plagiarism link). You may use any third party software, tools or packages, as long as their copyright does not prohibit their use for such purposes. If you have any form of question regarding this, please ask one of the lecturers, to avoid any misunderstanding.

Question 1 [2]

What is the difference between a decision tree and a regression tree?

Question 2 [3]

Suppose the AQR algorithm’s STAR procedure, in the Specialize complexes in STAR to exclude Eneg by... subroutine, has thus far produced the STAR: (rainy = yes cloudy = true), (temperature = 20 rainy = no) and determined EXTENSION to be (rainy = no), (cloudy = true), (temperature < 50). What would STAR be after the execution of the statement: Let \( \text{STAR} \) be the set: \( x \land y, x \in \text{STAR}, y \in \text{EXTENSION} \);

Question 3 [5]

Would K-means or K-medoids yield better generalization ability? Discuss your answer.
Question 4 [5]
Name and discuss two classes of clustering methods.

Question 5 [5]
Comment on the feasibility of using genetic programming to perform hierarchical clustering.

Question 6 [5]
In general, most clustering algorithms require that the number of clusters to be used be specified beforehand. This tends to be problematic as one does not always know a priori how many clusters will be required (especially for new and previously unused data sets). Discuss a technique to circumvent or alleviate this situation.