1 Objectives

This assignment aims to achieve the following general learning objectives:

- To gain experience with a real-world data set, in the context of data mining and exploratory data analysis;
- To gain practical experience in the application of various data cleaning techniques;
- To gain practical experience in a variety of exploratory data analysis techniques, as discussed in this course;
- To experiment with one of the classical data mining algorithms discussed in this course;
- To gain experience in formal, scientific report writing.

2 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.

3 Submission Instructions

You will have to write a report for this assignment. The report should be in standard PDF format, preferably compiled using \LaTeX. You will have to submit only the report. No additional files of any sort should be submitted. Failure to follow any of these instructions will result in a zero mark for the assignment.

Upload just your PDF file (named s99999999.pdf, where 99999999 is your student number), to the appropriate assignment upload on the course website. Multiple uploads are allowed, but only the last one will be marked. The deadline is Wednesday, 21 October 2015, at 23:00.

4 Data Set

The data set that you will be analysing for this assignment is a real-world data set produced by Statistics South Africa. The data set contains data related to the Victims of Crime Survey for 2013 and 2014 (available online at http://interactive.statssa.gov.za:8282/webview/), and contains census-like data about South African households and individuals, as well as their experience of, and opinions on crime. For this analysis, you will work with the data related to individuals. You will be responsible for retrieving the required data set files in an appropriate format, as well as any documentation related to the survey itself and what the various attributes in the data set represent. The focus of the knowledge that you extract is up to you — investigate the data set, and decide on an interesting topic that you would like to focus your knowledge extraction upon. Your investigations of the data set may reveal unexpectedly interesting facts, which may guide what your analysis chooses to address. You should try to take an objective view on the data, and disregard any preconceived ideas you have on the topic.
5 Analysis

It is your job to prepare the data set using whatever techniques are appropriate. You must do some initial analysis on the data set using whichever exploratory data analysis (EDA) techniques you find appropriate (you may use techniques discussed during the lectures, or appropriate techniques you find through other sources). You must address the following points during your analysis:

- You must appropriately clean the data set in whatever way is necessary before your data analysis can commence. You may have to perform analysis to determine what type of cleaning is necessary.

- You must analyse the data set using exploratory data analysis (EDA) techniques of your choosing. You will have to determine which techniques are appropriate for the data set, and which produce results that are the most informative. These techniques should involve data visualisation. You should include any relevant data visualisations as figures in your report.

- Use one of the classical data mining algorithms broadly discussed during this course (i.e. a rule induction algorithm, or a tree-building algorithm). You should report averaged results over at least thirty simulations performed on a cross-validated data set. Include standard deviations of your results. You should use these statistics to discuss the performance of the algorithm you have used.

- Your analysis may focus on any aspect of the data set. Ensure that your analysis is focused on this aspect, however. Do not perform too broad an analysis by focusing on several aspects, and end up not delivering interesting insight on any of them.

6 Report

You must write a report describing the analysis you performed in section 5. The report should be of an academic nature. This means that:

- The report’s structure should include all the aspects typically required of an academic paper (these include a title, abstract, introduction, methodology discussion, results, and conclusions).

- The tone of language should be formal and scientific, and must use correct spelling and grammar.

- Write as if your audience is familiar with the field of computer science, but not with artificial intelligence, data mining, or exploratory data analysis. Do not leave out any details. Explain every step you performed during your data preparation and analysis.

- Your report should be concise and to the point. Avoid discussing irrelevant details that don’t reveal anything interesting about the data set, your data preparation, or your analysis. However, make sure that you describe all the details of your data analysis. This is especially the case in relation to data cleaning and preprocessing, where you should describe both the steps that you performed, and the ones you did not perform. For the data preparation tasks that were not performed, you should explain why you performed no processing. This is the case even if it seems obvious to you.

- An adequate background discussion should be provided. This means that you must describe the data set in broad terms. You must also discuss every technique that you use, and provide a reference to published sources that describe the techniques.

- Adequate references must be provided. This means that you must adequately cite all techniques that you discuss, and all the reference details must be correct. It is not sufficient to simply cite the course slides or Wikipedia.

- Any figures or graphs included should be clear and of a professional standard. Include all information needed to interpret a figure (for example a legend and axis labels for a graph). Note that it is insufficient to include screen shots taken directly from analysis software. Every figure or graph requires an explanatory caption. You must provide a complete discussion of each figure and table in the text. If I cannot understand what your figures are illustrating, I will not be able to give you credit for the analysis.

It is recommended that you consult several existing conference and journal papers, as a guide to the type of style you should adopt. There are many such sources freely available online. It is very strongly recommended that you use the \LaTeX template for IEEE conference papers (available as \texttt{LatexSample.zip} in the folder for assignments, on the course website), to typeset your paper. A typical conference-length paper is approximately eight pages in length, and you should aim for roughly this length (it is safe to extend beyond this limit due to large figures, but avoid an excessively long discussion — keep your report focused).
7 Marking

The following general breakdown will be used during the assessment of this assignment:

<table>
<thead>
<tr>
<th>Category</th>
<th>Mark Allocation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing style</td>
<td>10 marks</td>
</tr>
<tr>
<td>Report structure</td>
<td>10 marks</td>
</tr>
<tr>
<td>References</td>
<td>10 marks</td>
</tr>
<tr>
<td>Background information</td>
<td>20 marks</td>
</tr>
<tr>
<td>Data cleaning discussion</td>
<td>20 marks</td>
</tr>
<tr>
<td>Exploratory analysis results</td>
<td>10 marks</td>
</tr>
<tr>
<td>Rule extraction results</td>
<td>10 marks</td>
</tr>
<tr>
<td>Conclusions</td>
<td>10 marks</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>100 marks</strong></td>
</tr>
</tbody>
</table>