INFO411: Data Mining and Knowledge Discovery

Project 10

Instructions:

This task is a real-world data mining problem. You are required to prepare a set of presentation slides which must include (1) the full name and student number of each student in the group, the contribution (in percent) of each group member, (2) your proposed data mining approach and methodology; (3) the strengths and weaknesses of your proposed approach; (4) the performance measures that can evaluate your data mining results; (5) the results and a brief discussion. Below is the recommended structure of your slides:

- Introduction (define the problem and the goal)
- Methods (propose approaches, and discuss their strengths and weaknesses)
- Results (Figures and tables of data analysis)
- Discussion (discovered knowledge from data mining)

Task: Convict Transportation

Background:
An estimated 160,000 convicts were transported to Australia during its colonial era. These include prisoners sent to New SouthWales, Van Diemens Land (Tasmania), Moreton Bay (Brisbane), Port Phillip. Western Australia and Norfolk Island. The database available at https://data.gov.au/dataset/ds-qld-458eb59f-e5f1-466f-925b-9dbcebb4f073/details?q=british%20convict
includes over 123,000 out of these convicts, over the period of 1787 to 1867. Some ships which were bound for Gibraltar are also recorded. You are required to restrict your analysis to the last 40 years of transportation, i.e. the period 1828-1867. The dataset includes a lot of text entries, most of which are not relevant to this data mining task. A particular feature of the sentence durations is that some are “life” sentences rather than a number of years. So you will need to consider how to handle this mixed measurement type.

Requirements:

1. Propose two different prediction algorithms to predict the duration of the sentence as a function of the place of trial, the place of arrival, and the date of departure.
2. Also, propose a classification algorithms to predict whether the convict received a ticket of leave as a function of the place of trial, the place of arrival, and the date of departure, and the sentence.
3. Present details of data pre-processing. This could include some merging of categories, if this is considered to be appropriate.
4. Explain why weak prediction or classification performance might still be of potential historical interest in this particular application.
5. Discuss the strengths and weaknesses of proposed models, and present your preferred models.
6. Present performance measures of your classification results.
7. Discuss whether or not the duration of sentences tended to change over time.