INFO411: Data Mining and Knowledge Discovery

Project 17

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: Pet Registration and Trends

Background:
Under Domestic Animals Act 1994, all domestic cats and dogs in Victoria over three months old must be registered with the local council. The database available at

(`https://www.data.gov.au/dataset/dandenong-registered-cats`) and (`https://www.data.gov.au/dataset/dandenong-registered-dogs`)

contains information about the suburb and postal code of each registered animal, the primary breed, primary colour, its month of birth, and whether it has been desexed.

There are over 200 dog and 50 cat breeds listed (though some are crosses) and a comparable number of colours, though, again, some are mixes. You may need to think about how to handle such a large number of levels of a factor. You may also want to look up the demographics of the suburbs in question over time.

Requirements:

1. Download the latest version of the datasets from the websites above.
2. Explore the relationship between the suburb (and its demographics) and the popularity of the breeds and colours of cats and dogs registered there, as well as their popularity over time (as measured by year and month of birth) using visualisation, numerical summaries, and models, and present your results.
3. Is there evidence of changes in relative popularity of cats and dogs between pets born up to 2014 and those born later? Does the change between these two periods differ by postcode?
4. Propose two different models to predict whether an animal will be desexed based on other information available (species, location, when born, etc.).
5. Present details of data pre-processing. This could include some merging of categories, representing mixes differently, and other techniques.
6. Discuss the strengths and weaknesses of proposed models, and present your preferred model.
7. Present detailed numeric and graphical performance measures of your classification results.