COMP41680 Assignment 2

Deadline: Sunday 7th May 2023

Overview:

This assignment involves working with a dataset of bank customer satisfaction survey responses. The general objective here is to predict whether customers are satisfied with the bank based on their survey responses. The dataset consists of two different representations for the same set of customers:

1. **Tabular data** (*bank-tabular.csv*) which contains demographic information about each customer and ratings scores (1-5) relating to different aspects of their bank. A manually-labelled 'satisfied' variable indicates whether a customer is 'satisfied' or 'dissatisfied' with the bank.

2. **Text data** (*bank-comments.csv*) which contains short textual comments indicating customers' opinions regarding the bank.

Task 1. Data Preparation

- Download the data from the assignment from the link below. The ZIP file includes a README.TXT file which explains the data format:

  http://mlg.ucd.ie/modules/python/assign2/bank-data.zip

- Load the tabular data and applying appropriate preprocessing steps to address different data quality issues and to prepare it for analysis.

Task 2. Data Characterisations

- Analyse, characterise, and summarise the cleaned tabular dataset, using tables and visualisations where appropriate. This should include the analysis of customer demographic features and numeric ratings features. You should also explore temporal aspects of the data (e.g. whether customer satisfaction levels are changing over time).

Task 3. Tabular Data Classification

- Using the 'satisfied' variable in the data, explore the use of tabular data with different classifiers to automatically distinguish between "satisfied" and "unsatisfied" customer responses.
- Quantity the performance of the various classification models using an appropriate evaluation strategy. Report and discuss the evaluation results.

Task 4. Text Data Classification

- Load the text comments data and integrate this with the existing tabular data, applying any necessary preprocessing steps.

- Using the ‘satisfied’ variable in the data, explore the use of the text data with different classifiers to automatically distinguish between “satisfied” and “unsatisfied” customer responses.
- Quantity the performance of the various classification models using an appropriate evaluation strategy. Report and discuss the evaluation results.

Task 5. Conclusions

- Discuss and compare the overall performance of the two different data representations (i.e. tabular and text) for customer satisfaction classification.
- At the end of your notebook, summarise any insights which you gained from your analysis of the data, discuss the challenges faced, and suggest ideas for further analysis/classification which could be performed on the data.

Guidelines:

- The assignment should be completed individually. All submissions will be subject to plagiarism checking. Any evidence of plagiarism will result in a 0 grade.
- The grade awarded will depend on the complexity of the analysis and level of detail, i.e., data preprocessing, classifier evaluation and comparison etc.
- Submit your assignment via the COMP41680 Brightspace page. Your submission should be in the form of a single ZIP file containing your notebook (i.e. IPYNB file).
- Hard deadline: Submit by end of 7th May 2023.
  Penalties will apply for late submissions:
    - 1-5 days late: 1 grade point deduction, e.g. B to B-
    - 6-10 days late: 2 grade point deduction, e.g. B to C+
    - Assignments will not be accepted later than 10 days without Extenuating Circumstances formally approved by UCD.