RFN1 — RFN1 TASK 2: CLUSTERING TECHNIQUES

MACHINE LEARNING — D603
PRFA — RFN1

COMPETENCIES

4163.1.2 : Recommends an Unsupervised Machine Learning Model

The learner recommends an unsupervised machine learning model based on a comparison of model performance given a business problem.

INTRODUCTION

In this task, you will act as an analyst and create a data mining report. In doing so, you must select one of the data dictionary and dataset files to use for your report from the following options:
churn_clean.csv and Churn Data Considerations and Dictionary.pdf
medical_clean.csv and Medical Data Considerations and Dictionary.pdf

You will use Python or R to analyze the given data and create a data mining report in a word processor (e.g., Microsoft Word). Throughout the submission, you must visually represent each step of your work and the findings of your data analysis.

SCENARIO

Scenario 1
One of the most critical factors in customer relationship management that directly affects a company's long-term profitability is understanding the customers. When a company understands its customers' characteristics, it is better able to target products and marketing campaigns for customers, resulting in better profits for the company in the long term.

You are an analyst for a telecommunications company that wants to better understand the characteristics of its customers. You have been asked to use clustering techniques to analyze customer data to identify groups of customers with similar characteristics, ultimately enabling better business and strategic decision-making.

Scenario 2
One of the most critical factors in patient relationship management that directly affects a hospital's long-term cost-effectiveness is understanding the patients and the condit

leading to hospital admissions. When a hospital understands its patients' characteristics, it is better able to target treatment to patients, resulting in a more effective cost of care for the hospital in the long term.

You are an analyst for a hospital that wants to better understand the characteristics of its patients. You have been asked to use clustering techniques to analyze patient data to identify groups of patients with similar characteristics, ultimately enabling better business and strategic decision-making for the hospital.

REQUIREMENTS

Your submission must represent your original work and understanding of the course material. Most performance assessment submissions are automatically scanned through the WGU similarity checker. Students are strongly encouraged to wait for the similarity report to generate after uploading their work and then review it to ensure Academic Authenticity guidelines are met before submitting the file for evaluation. See Understanding Similarity Reports for more information.

Grammarly Note:
Professional Communication will be automatically assessed through Grammarly for Education in most performance assessments before a student submits work for evaluation. Students are strongly encouraged to review the Grammarly for Education feedback prior to submitting work for evaluation, as the overall submission will not pass without this aspect passing. See Use Grammarly for Education Effectively for more information.

Microsoft Files Note:
Write your paper in Microsoft Word (.doc or .docx) unless another Microsoft product, or pdf, is specified in the task directions. Tasks may not be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc. All supporting documentation, such as screenshots and proof of experience, should be collected in a pdf file and submitted separately from the main file. For more information, please see Computer System and Technology Requirements.

You must use the rubric to direct the creation of your submission because it provides detailed criteria that will be used to evaluate your work. Each requirement below may be evaluated by more than one rubric aspect. The rubric aspect titles may contain hyperlinks to relevant portions of the course.

A. Create your subgroup and project in GitLab using the provided web link by doing the following:
- Clone the project to the IDE.
- Commit with a message and push when you complete each requirement listed in parts D and E.

  Note: You may commit and push whenever you want to back up your changes, even if a requirement is not yet complete.

- Submit a copy of the GitLab repository URL in the "Comments to Evaluator" section when you submit this assessment.
- Submit a copy of the repository branch history retrieved from your repository, which must include the commit messages and dates.

B. Describe the purpose of this data mining report by doing the following:
1. Propose **one** question relevant to a real-world organizational situation that you will answer using **one** of the following clustering techniques:
   - *k*-means
   - hierarchical

Note: The implementation of k-means clustering requires the use of continuous variables only.

2. Define **one** goal of the data analysis. Ensure your goal is reasonable within the scope of the selected scenario and is represented in the available data.

C. Explain the reasons for your chosen clustering technique from part B1 by doing the following:
   1. Explain how the clustering technique you chose analyzes the selected dataset. Include expected outcomes.
   2. Summarize **one** assumption of the clustering technique.
   3. List the packages or libraries you have chosen for Python or R, and justify how *each* item on the list supports the analysis.

D. Perform data preparation for the chosen dataset by doing the following:
   1. Describe **one** data preprocessing goal relevant to the clustering technique from part B1.
   2. Identify the initial dataset variables you will use to perform the analysis for the clustering question from part B1, and label *each* as continuous or categorical.
   3. Explain *each* of the steps used to prepare the data for the analysis. Identify the code segment for *each* step.
   4. Provide a copy of the cleaned dataset.

E. Perform the data analysis and report on the results by doing the following:
   1. Determine the optimal number of clusters in the dataset, and describe the method used to determine this number.

F. Summarize your data analysis by doing the following:
   1. Visualize the clusters and explain the quality of the clusters created. Include a screenshot of the cluster visualizations.
   2. Discuss the results and implications of your clustering analysis.
   3. Discuss **one** limitation of your data analysis.
   4. Recommend a course of action for the real-world organizational situation from part B1 based on the results and implications discussed in part F2.

G. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the programming environment.

Note: The audiovisual recording should feature you visibly presenting the material (i.e., not in voiceover or embedded video) and should simultaneously capture both you and your multimedia presentation.

Note: For instructions on how to access and use Panopto, use the "Panopto How-To Videos" web link provided below. To access Panopto's website, navigate to the web link titled "Panopto Access," and then choose to log in using the "WGU" option. If prompted, log in using your WGU student portal credentials, and then it will forward you to Panopto's website.

To submit your recording, upload it to the Panopto drop box titled "Task 2: Clustering Techniques – RFN1 | D603." Once the recording has been uploaded and processed in Panopto's system, retrieve the URL of the recording from Panopto and copy and paste it into the Links option. Upload the remaining task requirements using the Attachments option.

H. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure the web sources are reliable.

I. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.

J. Demonstrate professional communication in the content and presentation of your submission.

File Restrictions