RFN1 — RFN1 TASK 1: CLASSIFICATION DATA MINING MODELS

MACHINE LEARNING — D603

PRFA — RFN1

Preparation Task Overview Submissions Evaluation Report

COMPETENCIES

4163.1.1 : Recommends a Supervised Machine Learning Model

The learner recommends a supervised 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 should also refer to the data dictionary file for your chosen dataset from the provided options. 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.

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 classification methods:
   - Random forest
   - AdaBoost
   - Gradient boost
2. Define **one** goal of the data analysis. Ensure that your goal is reasonable within the scope of the scenario and is represented in the available data.

C. Explain the reasons for your chosen classification method from part B1 by doing the following:
1. Explain how the classification method you chose analyzes the selected dataset. Include expected outcomes.
2. 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 classification method from part B1.
2. Identify the initial dataset variables that you will use to perform the analysis for the classification question from part B1, and classify *each* variable 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. Split the data into training, validation, and test datasets and provide the file(s).
2. Create an initial model using the training dataset and provide a screenshot of the following metrics:
   - accuracy
   - precision
   - recall
   - F1 score
   - AUC-ROC
   - confusion matrix

3. Perform hyperparameter tuning on the validation dataset using k-fold cross validation to find the optimized model. Provide the following in the submission:
    - identification of which hyperparameters were selected for tuning
    - justification of the selection of these hyperparameters
    - screenshot of the best hyperparameters
4. Use the optimized model identified in part E3 to make predictions using the test dataset and provide a screenshot of the following metrics:
    - accuracy
    - precision
    - recall
    - F1 score
    - AUC-ROC
    - confusion matrix

F. Summarize your data analysis by doing the following:

1. Compare and discuss the metrics of accuracy, precision, recall, F1 score, and AUC-ROC from the use of the optimized model on the test dataset and the initial model on the training dataset to evaluate the performance of the optimized model.
2. Discuss the results and implications of your classification 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 your 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 1: Classification Data Mining Models – 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

File name may contain only letters, numbers, spaces, and these symbols: ! - _ . * ' ( )
File size limit: 200 MB
File types allowed: doc, docx, rtf, xls, xlsx, ppt, pptx, odt, pdf, csv, txt, qt, mov, mpg, avi, mp3, wav, mp4, wma, flv, asf, mpeg, wmv, m4v, svg, tif, tiff, jpeg, jpg, gif, png, zip, rar, tar, 7z