NVM4 — NVM4 TASK 1: CLASSIFICATION ANALYSIS

DATA MINING I — D209

PRFA — NVM4

COMPETENCIES

4030.6.1 : Classification Data Mining Models

The graduate applies observations to appropriate classes and categories using classification models.

4030.6.3 : Data Mining Model Performance

The graduate evaluates data mining model performance for precision, accuracy, and model comparison.

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 data set files to use for your report from the following link: Data Sets and Associated Data Dictionaries.

You should also refer to the data dictionary file for your chosen data set from the provided link. 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.

Note: All algorithms and visual representations used need to be captured either in tables or as screenshots added into the submitted document. A separate Microsoft Excel (.xls or .xlsx) document of the cleaned data should be submitted along with the written aspects of the data mining report.

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.

Part I: Research Question

A. 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:
      - *k*-nearest neighbor (KNN)
      - Naive Bayes

   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.

Part II: Method Justification

B. Explain the reasons for your chosen classification method from part A1 by doing the following:
   1. Explain how the classification method you chose analyzes the selected data set. Include expected outcomes.
   2. Summarize **one** assumption of the chosen classification method.
   3. List the packages or libraries you have chosen for Python or R and justify how *each* item on the list supports the analysis.

Part III: Data Preparation

C. Perform data preparation for the chosen data set by doing the following:
   1. Describe **one** data preprocessing goal relevant to the classification method from part A1.
   2. Identify the initial data set variables that you will use to perform the analysis for the classification question from part A1 and classify *each* variable as numeric 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 data set.

Part IV: Analysis

D. Perform the data analysis and report on the results by doing the following:
   1. Split the data into training and test data sets and provide the file(s).
   2. Describe the analysis technique you used to appropriately analyze the data. Include screenshots of the intermediate calculations you performed.
   3. Provide the code used to perform the classification analysis from part D2.

Part V: Data Summary and Implications

E. Summarize your data analysis by doing the following:
   1. Explain the accuracy and the area under the curve (AUC) of your classification 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 A1 based on your results and implications discussed in part E2.

Part VI: Demonstration

F. 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 "Data Mining I - NVMx | D209 (student creators) [assignments]." 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.

G. Acknowledge web sources, using in-text citations and references, for segments of third-party code or data used to support the analysis. Be sure the web sources are reliable.

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

I. 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

RUBRIC

A1:PROPOSAL OF QUESTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not propose 1 question.	The submission proposes 1 question that is not relevant to a real-world organizational situation. Or the proposal does not include 1 of the given classification methods.	The submission proposes 1 question that is relevant to a real-world organizational situation, and the proposal includes 1 of the given classification methods.

A2:DEFINED GOAL