NVM2 — NVM2 TASK 1: CLASSIFICATION ANALYSIS

DATA MINING I — D209
PRFA — NVM2

TASK OVERVIEW SUBMISSIONS EVALUATION REPORT

COMPETENCIES

4030.06.1 : Classification Data Mining Models
The graduate applies observations to appropriate classes and categories using classification models.

4030.06.3 : 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 be your original work. No more than a combined total of 30% of the submission and no more than a 10% match to any one individual source can be directly quoted or closely paraphrased from sources, even if cited correctly. The originality report that is provided when you submit your task can be used as a guide.

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.

*Tasks may **not** be submitted as cloud links, such as links to Google Docs, Google Slides, OneDrive, etc., unless specified in the task requirements. All other submissions must be file types that are uploaded and submitted as attachments (e.g., .docx, .pdf, .ppt).*

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 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 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 – NVM2." 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. Record the web sources used to acquire data or segments of third-party code to support the analysis. Ensure 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, 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

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not define 1 goal for data analysis.	The submission defines 1 goal for data analysis, but the goal is not reasonable, is not within the scope of the scenario, or is not	The submission defines 1 reasonable goal for data analysis that is within the scope of the scenario and is represented in the available data.

represented in the available data.

B1:EXPLANATION OF CLASSIFICATION METHOD

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not explain how the chosen classification method analyzes the selected data set.	The submission does not logically explain how the chosen classification method analyzes the selected data set, or the explanation includes inaccurate expected outcomes.	The submission logically explains how the chosen classification method analyzes the selected data set and includes accurate expected outcomes.

B2:SUMMARY OF METHOD ASSUMPTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not summarize 1 assumption of the chosen classification method.	The submission inadequately summarizes 1 assumption of the chosen classification method.	The submission adequately summarizes 1 assumption of the chosen classification method.

B3:PACKAGES OR LIBRARIES LIST

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not list the packages or libraries chosen for Python or R.	The submission lists the packages or libraries chosen for Python or R but does not justify how 1 or more items on the list support the analysis.	The submission lists the packages or libraries chosen for Python or R and justifies how *each* item on the list supports the analysis.

C1:DATA PREPROCESSING

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not describe 1 data preprocessing goal.	The submission describes 1 data preprocessing goal, but it is not relevant to the classification method from part A1.	The submission describes 1 data preprocessing goal that is relevant to the classification method from part A1.

C2:DATA SET VARIABLES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not identify *any* data set variables used to perform the analysis for the classification question from part A1 or does not classify the variables as continuous or categorical.	The submission identifies the data set variables used to perform the analysis for the classification question from part A1, but the submission inaccurately classifies 1 or more variables as continuous or categorical.	The submission identifies the data set variables used to perform the analysis for the classification question from part A1, and the submission accurately classifies *each* variable as continuous or categorical.

C3:STEPS FOR ANALYSIS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not explain *each* step used to prepare the data for the analysis, or the submission does not identify the code segment for *each* step.	The submission inaccurately explains 1 or more steps used to prepare the data for analysis, or the submission identifies an inaccurate code segment for 1 or more steps.	The submission accurately explains *each* step used to prepare the data for analysis, and the submission identifies an accurate code segment for *each* step.

C4:CLEANED DATA SET

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include a copy of the cleaned data set	The submission includes a copy of the cleaned data set, but the data set is inaccurate.	The submission includes an accurate copy of the cleaned data set.

D1:SPLITTING THE DATA

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide the training and test data set file(s).	The submission provides training and test data sets, but the split is not reasonably proportioned.	The submission provides reasonably proportioned training and test data sets.

D2:OUTPUT AND INTERMEDIATE CALCULATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT

The submission does not describe the analysis technique used to analyze the data, or it does not include screenshots of the intermediate calculations performed.	The submission inaccurately describes the analysis technique used to appropriately analyze the data, or the submission includes screenshots of the intermediate calculations performed but they are inaccurate.	The submission accurately describes the analysis technique used to appropriately analyze the data, and the submission includes accurate screenshots of the intermediate calculations performed.

D3:CODE EXECUTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide the code used to perform the classification analysis from part D2.	The submission provides the code used to perform the classification analysis from part D2, but 1 or more errors are evident during the execution of the code.	The submission provides the code used to perform the classification analysis from part D2 and the code executes without errors.

E1:ACCURACY AND AUC

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not explain the accuracy or the AUC of the classification model.	The submission does not logically explain the accuracy or the AUC of the classification model.	The submission logically explains *both* the accuracy and the AUC of the classification model.

E2:RESULTS AND IMPLICATIONS

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss *both* the results and implications of the classification analysis.	The submission discusses *both* the results and implications of the classification analysis, but the discussion is inadequate.	The submission adequately discusses *both* the results and implications of the classification analysis.

E3:LIMITATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not discuss 1 limitation of the data analysis.	The submission discusses 1 limitation of the data analysis but lacks adequate detail or is illogical.	The submission logically discusses 1 limitation of the data analysis with adequate detail.

E4:COURSE OF ACTION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not recommend a course of action for the real-world organizational situation from part A1	The submission does not recommend a reasonable course of action for the real-world organizational situation from part A1, or the course of action is not based on the results and implications discussed in part E2.	The submission recommends a reasonable course of action for the real-world organizational situation from part A1 based on the results and implications discussed in part E2.

F:PANOPTO RECORDING

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not provide a Panopto video recording.	The submission provides a Panopto video recording, but it does not include a demonstration of the functionality of the code used for the analysis or a summary of the programming environment or *both*.	The submission provides 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.

G:SOURCES FOR THIRD-PARTY CODE

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not record web sources used to acquire data or segments of third-party code.	The submission records 1 or more unreliable web sources used to acquire data or segments of third-party code.	The submission records *all* web sources used to acquire data or segments of third-party code, and the web sources are reliable.

H:SOURCES

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
The submission does not include both in-text citations and a reference list for sources that are quoted, paraphrased, or summarized.	The submission includes in-text citations for sources that are quoted, paraphrased, or summarized and a reference list; however, the citations or reference list is incomplete or inaccurate.	The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available.

I:PROFESSIONAL COMMUNICATION

NOT EVIDENT	APPROACHING COMPETENCE	COMPETENT
Content is unstructured, is disjointed, or contains pervasive errors in mechanics, usage, or grammar. Vocabulary or tone is unprofessional or distracts from the topic.	Content is poorly organized, is difficult to follow, or contains errors in mechanics, usage, or grammar that cause confusion. Terminology is misused or ineffective.	Content reflects attention to detail, is organized, and focuses on the main ideas as prescribed in the task or chosen by the candidate. Terminology is pertinent, is used correctly, and effectively conveys the intended meaning. Mechanics, usage, and grammar promote accurate interpretation and understanding.

WEB LINKS

Data Sets and Associated Data Dictionaries
If you have trouble with the link, copy and paste the link directly into your web browser.

Panopto Access
Sign in using the "WGU" option. If prompted, log in with your WGU student portal credentials, which should forward you to Panopto's website. If you have any problems accessing Panopto, please contact Assessment Services at assessmentservices@wgu.edu. It may take up to two business days to receive your WGU Panopto recording permissions once you have begun the course.

Panopto How-To Videos