COMPETENCIES

4030.6.2 : **Predictive Data Mining Models**
The graduate implements prediction data mining models to find hard-to-spot relationships among variables.

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 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 similarity 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 prediction methods:

- decision trees
- random forests
- advanced regression (i.e., lasso or ridge regression)

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 prediction method from part A1 by doing the following:
   1. Explain how the prediction method you chose analyzes the selected data set. Include expected outcomes.
   2. Summarize **one** assumption of the chosen prediction 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 prediction method from part A1.
   2. Identify the initial data set variables that you will use to perform the analysis for the prediction question from part A1 and group *each* variable as numeric or categorical.
   3. Explain 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 prediction 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 mean squared error (MSE) of your prediction model.
   2. Discuss the results and implications of your prediction 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.